Learning to Adapt in Dialogue Systems: Data-driven Models for Personality Recognition and Generation

François Mairesse

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Department of Computer Science
University of Sheffield, United Kingdom

February 2008

Abstract

Dialogue systems are artefacts that converse with human users in order to achieve some task. Each step of the dialogue requires understanding the user's input, deciding on what to reply, and generating an output utterance. Although there are many ways to express any given content, most dialogue systems do not take linguistic variation into account in both the understanding and generation phases, i.e. the user's linguistic style is typically ignored, and the style conveyed by the system is chosen once for all interactions at development time. We believe that modelling linguistic variation can greatly improve the interaction in dialogue systems, such as in intelligent tutoring systems, video games, or information retrieval systems, which all require specific linguistic styles. Previous work has shown that linguistic style affects many aspects of users' perceptions, even when the dialogue is task-oriented. Moreover, users attribute a consistent personality to machines, even when exposed to a limited set of cues, thus dialogue systems manifest personality whether designed into the system or not. Over the past few years, psychologists have identified the main dimensions of individual differences in human behaviour: the Big Five personality traits. We hypothesise that the Big Five provide a useful computational framework for modelling important aspects of linguistic variation. This thesis first explores the possibility of recognising the user's personality using data-driven models trained on essays and conversational data. We then test whether it is possible to generate language varying consistently along each personality dimension in the information presentation domain. We present PERSONAGE: a language generator modelling findings from psychological studies to project various personality traits. We use PERSONAGE to compare various generation paradigms: (1) rule-based generation, (2) overgenerate and select and (3) generation using parameter estimation models—a novel approach that learns to produce recognisable variation along meaningful stylistic dimensions without the computational cost incurred by overgeneration techniques. We also present the first human evaluation of a data-driven generation method that projects multiple stylistic dimensions simultaneously and on a continuous scale.

Acknowledgements

Here I would like to express my sincere gratitude to the many people who have contributed to making this thesis an enjoyable and rewarding experience.

First of all, I would like to thank Lyn Walker for her enthusiastic supervision, and for making me discover the field of computational linguistics, as well as research methods in general. This thesis could not have existed without her. I am also grateful to Roger Moore for our fruitful conversations, as well as to Rob Gaizauskas for his valuable advice. On a different continent, I would like to thank Matthias Mehl, James Pennebaker and Cindy Chung for their collaboration, and for introducing me to the world of psychology research.

On a more personal note, I would like to thank my family—Annick, Pierre, Cécile and Jérémie—who I should have seen much more often, as well as Matt Gibson, Sarah Creer and Helen Cross for their humour, kindness and friendship throughout these three years, and the Cobden View staff for making Tuesday the quiz day. Many thanks to John Allan, Joe Polifroni, Jonathan Laidler and Nathaniel King for their conversations, lunches and the occasional pint. I am also grateful to the departmental staff and support for making the department work, and especially to Gillian Callaghan for her northern humour. I also had the pleasure of living with many people—with whom I hope I have shared a bit more than a house—including Rachel Fairlamb, Jonathan Chambers, Catherine Quigley, Dean Young, Jonathon Rudkin, Dave Robinson, Hector Marin Reyes, Bouka Maamra, Yannis Balaskas, Sotiris Saravelos, Tudor Grecu and Daniele Musumeci.

While Sheffield has been my new home for the past three years, I must also thank the people who made me want to go back to Belgium (and not forget about strong fizzy beers): François Delfosse, Simon Dubois, Geoffroy Tassenoy, Guillaume De Bo, Benoît Roland, Lionel Lawson, Ludovic Lecocq, Fabrice Quinet, Quentin Vroye, Rebecca Deraeck, Nathalie Hargot, Maxime Melchior, Benjamin Cogels, Gérard Paquet and many others.

Lastly, and most importantly, I would like to thank Véronique Lefebvre for her constant affection and smile.

Table of Contents

Chapter 1

Introduction

As machines increase in complexity, so do the communication mechanisms required to control them. This complexity typically requires specialised training, putting the burden on the human operator to adapt to its tool. Recently, the evolution of computing has reached a point where machines are capable of understanding and imitating natural human behaviour. This implies that the future is likely to see a shift of the burden from the user to the machine, as computers are expected to communicate using a complex, ill-defined protocol that is the most familiar to the user: *natural language*.

As a result, a new type of computer system has emerged—referred to as *dialogue systems*—whose purpose is to communicate with human users using natural language. These systems typically need to (1) understand the user's utterance, (2) decide on what action to take (dialogue management), and (3) produce a corresponding output utterance, while making sure the dialogue is natural enough to be successful. Figure 1.1 illustrates the resulting information flow during a conversation with the user.

While language is sometimes seen as a protocol for exchanging information explicitly, it is more than that: language provides cues which are used by the hearer to infer properties about the speaker's state or intentions. The whole field of *pragmatics* is dedicated to the study of information that is not conveyed explicitly through language. Furthermore, speakers also use linguistic cues—consciously or not—to project their emotions, moods and personality, regardless of the information they communicate. The resulting individual differences in language are referred to as *linguistic variation*.

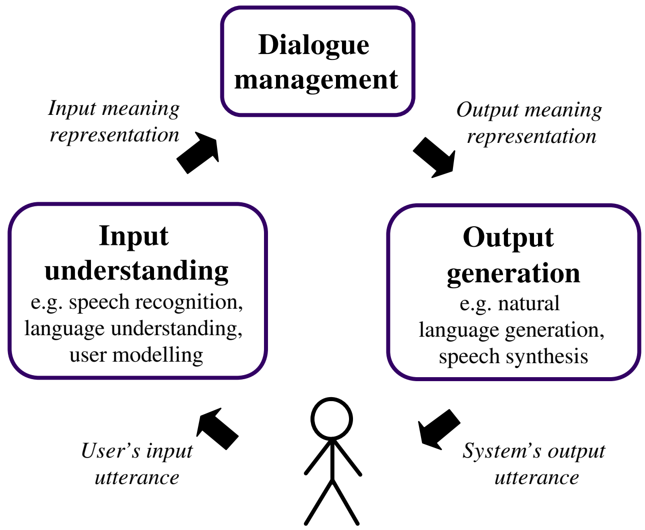


Figure 1.1: High-level architecture of a dialogue system.

1.1 Modelling the user's linguistic variation

Most current dialogue systems do not take linguistic variation into account. During the language understanding phase, keyword spotting techniques are generally used to associate the user's utterance to a meaning representation, without affecting the model of the user's state. For example, let us assume that an anxious user tries to book a flight through the phone. Even if the system understands his or her request, it will treat the user as any other, and ignore any cue revealing the user's anxiety. Given a high recognition confidence score, the system is likely to choose not to confirm the user's request, which can then lead the anxious user to request a confirmation, thus adding complexity to the dialogue as well as potential sources of error. If the system had detected the user's anxiety, it could have acted accordingly, e.g. by enforcing implicit confirmations. Even more so, the system could have changed its linguistic output to produce more reassuring utterances. This leads us to another essential aspect of adaptation in dialogue: the use of appropriate linguistic variation in the system's output.

1.2 Generating linguistic variation to the user

Computers are generally associated with monotonic, formal and repetitive language, as most commercial dialogue systems output generic template utterances to the user. The flexibility is usually limited to the insertion of variables within these templates, such as the variables AIRPORT, DATE and TIME in *'your flight will depart from AIRPORT on DATE, at TIME'*. Additional flexibility can be provided using a *natural language generator*, which is a software component that is dedicated to the generation of comprehensible language that conveys the system's communicative goal, e.g. the confirmation of the user's flight.

Although a large body of work has focused on natural language generation [Reiter and Dale, 2000], most generators aim at producing a single, coherent output given a specific communicative goal, thus implicitly ignoring the effect of linguistic variation. The example of the anxious speaker in the previous section suggests a utility for controlling this variation, in order to improve the user's experience as well as the dialogue's efficiency.

This chapter provides additional motivation for modelling linguistic variation in dialogue, by describing various findings related to this issue as well as potential applications. But first of all, what should the dimensions of linguistic variation be? The next section addresses this question by presenting some of the most important variation dimensions.

1.3 Dimensions of linguistic variation

Whenever we produce an utterance, we make many consistent conscious and subconscious choices in order to convey the information in the desired way. Why don't we only express basic facts in the most unambiguous way? That is just what computers do when they communicate between themselves, and it seems to be the most efficient way to accomplish a particular task. A possible reason is that speakers try to satisfy multiple goals simultaneously, such as communicating information, minimising their vocal effort, and projecting a specific image to the hearer. Individual speakers value these goals differently, resulting in different linguistic styles.

1.3.1 Definitions of linguistic style

Intuitively, linguistic style refers to a specific point within the space of all possible linguistic variation. More formally, Bell [1997] defines it as 'the range of variation within the speech of an individual speaker', while Labov [1984] proposes a more general definition:

> By 'style...' we mean to include any consistent [set of] linguistic forms used by a speaker, qualitative or quantitative, that can be associated with a [set of] topics, participants, channel, or the broader social context.

From these definitions, it seems that linguistic style can be considered as a temporary characteristic of a single speaker, as well as a long-lived property of a large group of the population. The most important element in Labov's definition is probably the reference to *consistent* language forms, i.e. a set of linguistic cues that are repeated over time, and that can thus be measured.

These definitions do not say anything about the causes of linguistic variation. The next section therefore explores factors affecting linguistic style, in order to investigate what variation dimensions should be modelled within human-machine conversations.

1.3.2 Factors affecting linguistic style

Linguistic style is affected by many variables. Some can be consciously manipulated by the speaker based on the situation (e.g. formality), while others are typically acquired over a long period of time (or are innate) and more likely to be the result of early conditioning (e.g. dialects or personality). This section presents some of these factors.

1.3.2.1 Formality

Formality is one of the most studied dimensions in sociolinguistics. Labov [1984] relates formality to the speaker's level of attention towards language. He characterises it by a series of principles, which include:

> Principle of formality: Any systematic observation defines a formal context in which more than the minimal attention is paid to speech.
>
> Principle of attention: Styles may be ordered along a single dimension, measured by the amount of attention paid to speech.

Labov therefore considers formality as one of the most important dimensions in sociolinguistics. Some researchers also associate formality with the level of explicitness [Heylighen and Dewaele, 2002]. This dimension can thus be related to the speaker's (non-)reliance on the context. From a goal-oriented perspective, the level of formality is influenced by the speaker's goals regarding communication efficiency (e.g. unambiguity), adaptation to the hearer (e.g. a distant superior), as well as the image that the speaker wishes to convey (e.g. professionalism, respectfulness).

1.3.2.2 Politeness

Other human goals include the need for self-esteem and respect from others [Brown and Levinson, 1987]. These goals have an enormous effect on language. By trying to fulfill them, people tend to bias their messages in a way that computer systems do not. This phenomenon is referred to as *politeness*. Politeness theories suggest that the use of politeness is dependent on the social distance and the difference of power between conversational partners, as well as on the threat of the speaker's communicative act towards the hearer [Brown and Levinson, 1987].

1.3.2.3 Dialects and sociolects

While formality and politeness are controlled by the speaker based on his or her goals and environment, we now turn to variation dimensions that are more culture-dependent. These dimensions are typically used to analyse variation between large groups of speakers.

Among the factors affecting the speaker's linguistic style is geographical location, resulting in different *dialects*. The use of a dialect can reinforce the speaker's feeling of being part of a community. Dialectical variation is important, but it typically concerns a localised part of the population, which limits the utility of dialect modelling in dialogue systems.

A sociolect is a language characterising speakers from a specific social background or status. For example, markers of British English tend to indicate upper class when used in American English. While sociolects can indicate socio-economic class, they can also be representative of the speaker's gender, age and ethnic background.

1.3.2.4 Personality

While dialects and sociolects refer to patterns of language that are collective by nature, individual linguistic style can also be rooted in education as well as genetic factors. Different people behave in different ways irrespective of their social group, and some behavioural cues are consistent over time. As language can be seen as the main interface between a human being and its social environment, language should be assumed to be strongly affected by individual differences. We believe that such idiosyncracies can be categorised in terms of *personality traits*. While there exists different frameworks for analysing personality, the *Big Five* model has recently emerged as a standard in psychology [Norman, 1963]. The Big Five dimensions are:

- Extraversion vs. introversion (sociable, assertive, playful vs. reserved, shy)
- Emotional stability vs. neuroticism (calm, unemotional vs. insecure, anxious)
- Agreeableness (friendly, cooperative vs. antagonistic, faultfinding)
- Conscientiousness (self-disciplined, organised vs. inefficient, careless)
- Openness to experience (intellectual, creative vs. shallow, unimaginative)

These personality traits influence more temporary changes of behaviour, such as emotions [Watson and Clark, 1992], which again have a large influence on language. Personality also affects the way speakers tend to fulfill their individual goals, therefore it can be considered as the highest source of behaviour variation between people.

Although each variation dimension presented in this section has its specificities, we believe that principled work on modelling individual differences in dialogue should first focus on the most independent dimensions, as more specific behaviours can be approximated as a combination of them. We thus choose to model linguistic variation using personality dimensions, based on the Big Five framework. Personality provides the following advantages for modelling linguistic variation:

- By definition, personality traits represent the most fundamental dimensions of variation of human behaviour;

- Personality traits are well defined in the psychology literature;
- A large amount of research has studied the relation between personality traits and behavioural cues, including language;
- Specific personality traits can be approximated by combining major traits, e.g. the Big Five;
- Personality can be assessed easily using validated questionnaires;
- As personality is a permanent characteristic, personality judgements can be collected for many different language samples;
- Personality modelling has been shown to be an important factor for the design of dialogue systems (see Section 1.4).

As machines are programmed by humans, their personality is usually defined by the programmer once for all interactions at development time, whether or not he or she is aware of it. While not much attention is currently given to this aspect of human-computer interaction, the next section presents evidence for the need for personality modelling in dialogue systems.

1.4 Motivation for personality-based dialogue modelling

Being able to recognise the personality of the user as well as controlling the personality conveyed by the system can have many benefits.

As personality is usually considered as a fundamentally human concept, one might think that users of a machine will keep that in mind when interacting with it. Reeves and Nass [1996] suggest that it is not true; people tend to treat computers and other media as real people. Users associate a personality to the machine they interact with, and they even prefer working with machines exhibiting the same personality as themselves. This *similarity-attraction* effect suggests that there is a need for modelling both the *user's* and the *system's* personality, in order for dialogue systems to adapt to the user, like humans do [Byrne and Nelson, 1965, Funder and Sneed, 1993].

While our long-term objective is to use personality as a mediating layer for adapting to the user, each aspect of this dual task—i.e. personality recognition and generation—presents many other benefits.

1.4.1 Recognising the user's personality

Psychology research shows that personality traits influence many aspects of individual behaviour when performing specific tasks. For example, personality traits influence leadership ability [Hogan et al., 1994], attitude towards machines [Sigurdsson, 1991], general job performance and sales ability [Furnham et al., 1999], teacher effectiveness [Rushton et al., 1987], and academic ability and motivation [Furnham and Mitchell, 1991, Komarraju and Karau, 2005]. A system capable of recognising personality could thus adapt its behaviour based on the user's capabilities. For example, Sigurdsson [1991] shows that neurotics have more difficulties using computers, while Komarraju and Karau [2005] show that introverts are less engaged in learning. Based on this information, an intelligent tutoring system could improve its performance by providing more navigational support for neurotic users, while engaging more with introvert users.

Recent work in artificial intelligence explores methods for the automatic detection of other types of pragmatic variation in text and conversation, such as emotion [Oudeyer, 2002, Liscombe et al., 2003], deception [Newman et al., 2003, Enos et al., 2006, Graciarena et al., 2006, Hirschberg et al., 2005], mood [Mishne, 2005], dominance and leadership in meetings [Rienks and Heylen, 2006], point of view or subjectivity [Wilson et al., 2004, Wiebe et al., 2004, Wiebe and Riloff, 2005], and sentiment or opinion [Turney, 2002, Pang and Lee, 2005, Popescu and Etzioni, 2005]. In contrast with these pragmatic phenomena, which may be relatively contextualised or short-lived, personality is usually considered to be a long-term, stable aspect of individuals [Scherer, 2003]. However, there is evidence that personality affects these other aspects of linguistic production. For example, there are strong relations between the extraversion and conscientiousness traits and positive affect, and between neuroticism and disagreeableness and negative affect [Watson and Clark, 1992]. Outgoing and energetic people (i.e. extravert) are more successful at deception, while apprehensive (i.e. neurotic) individuals are not as successful [Riggio et al., 1988]. Concerning leader identification, Hogan et al. [1994] show that effective leaders are more extravert, conscientious and emotionally stable. Finally, Oberlander and Nowson [2006] suggest that opinion mining could benefit from personality information. Thus this evidence suggests that incorporating personality models into these other tasks may improve accuracy.

On a broader scale, personality recognition could also affect the way companies customise their products, as Reeves and Nass [1996] suggest that personality would be a good indicator for market segmentation, i.e. to tailor solutions to different customers in order to maximise their satisfaction.

1.4.2 Controlling the system's personality

Many psychological studies suggest that controlling the personality projected by computer interfaces would be beneficial for human-computer interaction.

The generation of more human-like language was shown to reduce the user's cognitive load, resulting in better task performance [Campana et al., 2004]. Reeves and Nass [1996] find that users prefer systems which exhibit any kind of personality rather than an inconsistent set of cues. The only way to avoid such inconsistencies is to take personality into account when designing the system. Ruttkay et al. [2004] suggest that personality is an important design variable for developing embodied conversational agents. Interestingly, Giles and Powesland [1975] have shown that impressions about a student's intelligence, enthusiasm and self-confidence are more dependent on language samples than on photographs or school work samples. This suggests that language is one of the most important indicators of personality. As dialogue systems often represent a specific company, it is therefore important that the system's personality—or *persona*—reflects the desired corporate image [Cohen et al., 2004]. We believe that the design and development of a persona would be greatly simplified if the designer could specify the personality of the system using a small number of high-level parameters. It is important to note that our objective is not to mislead the user into thinking that he or she is interacting with a human being. As well as raising ethical issues, doing so would increase the user's expectations to a level that cannot be met by machines, which is likely to result in poor user satisfaction.

Additionally, many dialogue applications would benefit from specific personality types, some of which are listed in Table 1.1. Rushton et al. [1987] find that extravert teachers are more effective, suggesting that an intelligent tutoring system should produce extravert language. Furnham et al. [1999] report that general job performance is associated with conscientiousness and emotional stability, while potency (extraversion) correlates positively with sales figures and superior ratings. The authors also show that impulsivity is a significant performance predictor of telesales employees selling insurance. These findings can guide dialogue system designers to optimise the personality conveyed by an automated sales agent. Furthermore, we hypothesise that a system gathering information from the user—e.g. a database modification interface—-should be receptive (introvert), agreeable and conscientious. Whenever time is crucial—e.g. when requesting stock quotes or emergency advice—information should be clear and concise. This suggests the need for a conscientious, introvert and non-agreeable operator, avoiding superfluous politeness forms.

Application	**System's personality**
Tutoring system	extravert agreeable conscientious
Telesales system	extravert (potent) impulsive
System learning from user	introvert agreeable conscientious
Financial information retrieval	conscientious not extravert not agreeable
Video game characters	any personality type
Interactive drama system	any personality type
Realistic training system	any personality type
Psychotherapy	any personality type

Table 1.1: Hypothesised optimal personality traits for various dialogue system applications.

Some applications require multiple personality types. Those include systems used in the entertainment industry, such as embodied conversational agents or video game characters. Interactive narrative systems—which automatically generate interactive stories with complex characters—currently require the creation of many handcrafted dialogues [Mateas, 2007]. This 'authoring bottleneck' could be partially resolved by using a parameterisable language generator conveying different personality traits. Training systems would also benefit from the projection of various personalities. Examples include systems training practitioners to interview anxious patients [Hubal et al., 2000], as well as systems training soldiers to gather information from uncooperative civilians through tactical questioning [Department of the Army, 2006]. Finally, a recent line of research has focused on the use of virtual environments for psychotherapy. For example, a patient scared of public speaking can learn to manage his or her fear by giving a talk in front of a virtual audience made of various types of characters [Slater et al., 2004].

1.5 Research hypotheses

While the previous sections suggest that personality modelling can enhance dialogue applications, this work investigates *how* to model both the personality of the user and the system. To clarify our objectives, it is important to define the research questions that this thesis will address, as well as the issues that are beyond the scope of this work.

This thesis investigates two major hypotheses about how individual differences can be modelled in dialogue systems:

Hypothesis 1 *Statistical models can learn to predict the personality of unseen individuals from conversational data, on a continuous scale.*

Hypothesis 2 *The personality conveyed by generated utterances can be controlled by a computational model, which can be either derived from psycholinguistic knowledge or learnt from personality-annotated linguistic data.*

As this thesis represents one of the first large bodies of work on personality recognition and generation through language, it aims at answering many research questions. While some of these questions have been investigated in previous work, they have not been studied in the context of a dialogue interaction.

- **Personality recognition:**
  - Can statistical models trained on spoken language annotated with judgements of personality successfully predict the personality of unseen individuals?

  - How does the personality assessment method—e.g. using the speaker's own judgement or observer reports—affects modelling accuracy?
  - What linguistic features are the most useful for personality recognition?
  - Should personality be modelled as a set of continuous or discrete dimensions?

- **Personality generation:**
  - How can personality markers identified in psychological studies be reproduced in a natural language generation system?
  - Can these personality markers be used to convey recognisable personality in a specific application domain?
  - Can the perception of multiple personality traits be controlled continuously using statistical models trained on personality judgements of generated utterances?
  - Can general-domain personality recognition models successfully predict the personality of generated utterances?
  - Can the system project a target personality through a single utterance, or is more information needed?
  - What generation decisions are the most useful for conveying a specific personality trait?
  - How do data-driven generation techniques compare to a rule-based approach purely based on findings from the psychology literature?

While each of these questions will be addressed in the following chapters, it is also important to specify the limitations of the scope of this thesis.

1.5.1 Boundaries

As the modelling of personality in dialogue is a vast topic, this work focuses on the presentation and evaluation of personality recognition and generation *methods*, without investigating whether personality modelling improves task performance in specific applications. We speculate that the adaptation can be informed by existing psychological studies, such as those presented in Section 1.4.2. This thesis thus focuses on personality modelling accuracy—rather than dialogue task performance—by assuming that personality traits represent a suitable layer of parameters for dialogue system adaptation.

Furthermore, the models presented in this thesis are not meant to reproduce the cognitive mechanisms found in the brain. Although the architecture of language generators presents similarities with models of human language production, the implementation proposed in this thesis does not attempt to imitate how personality affects the human language production process. Our objective is to propose a tractable implementation that reproduces the effect of personality on language, as perceived by an observer.

A final limitation is that the present work focuses on personality modelling at the *linguistic* level, leaving aside high-level dialogue strategy and acoustic modelling.[1] However, we believe that the methodology presented in this thesis can be applied to model non-verbal markers of personality as well, both in the recognition and generation phases.

1.6 Contributions and organisation of the thesis

The research questions are addressed by following a principled methodology, based on data-driven approaches to personality modelling. While the next chapter summarises the related work, the rest of the thesis consists of two main parts, investigating techniques for (1) recognising the personality of the user and (2) controlling the personality conveyed by the system's utterances.

The first part presents and evaluates data-driven models for personality recognition (Chapter 3). The models are trained on personality judgements of written and spoken language, as well as on ratings of self-reported personality. As it is not clear whether personality is best modelled as a continuous variable, we successively treat the personality recognition task as a classification, regression and ranking problem. We find that recognition models outperform the baseline on unseen individuals for each trait of the Big Five model. Additionally, results show that observed personality is easier to model than self-assessed personality.

[1] This is only valid for the generation phase, as prosodic cues are included in personality recognition models in Chapter 3.

The second—and most important—part of this thesis investigates various methods for projecting personality using a natural language generation (NLG) system. For each of the Big Five traits, Chapter 4 reviews findings from psychological studies and systematically maps them to generation parameters within the standard NLG architecture. The resulting mappings offer a principled approach to the generation of personality markers, as they represent our hypotheses about how each finding can be reproduced in a computational artefact to convey personality in a new dialogue domain. While the parameter mappings are application-independent, Chapter 5 presents a detailed implementation of each parameter in PERSONAGE, a psychologically-informed generator that produces restaurant recommendations. This generator is used as a building block for comparing various generation paradigms in the following chapters. The PERSONAGE-RB rule-based generator uses predefined parameter settings derived from psychology findings to project extreme personality traits within the Big Five model. An evaluation of PERSONAGE-RB is presented in Chapter 6, showing that human judges recognise the personality of its output utterances for each end of the Big Five dimensions.

While PERSONAGE-RB relies on hypotheses derived from the psychology literature, Chapter 7 presents a 'bottom-up' correlational analysis of the linguistic features that influence the personality perceived by the judges. This analysis provides an implicit evaluation of the hypotheses made in Chapter 4, i.e. testing whether or not the psychology findings generalise to a specific application domain. Additionally, Chapter 7 analyses whether PERSONAGE can generate utterances covering the full range of the Big Five personality scales, by randomly varying its parameter values. While PERSONAGE-RB can only target a discrete set of traits, Chapter 8 presents a data-driven generation technique based on the 'overgenerate and select' paradigm, that can target any arbitrary personality value along the Big Five scales. This technique is implemented in the PERSONAGE-OS data-driven generator, which uses PERSONAGE to randomly generate many utterances, and then selects the utterance projecting the desired personality using a statistical regression model trained on human personality judgements. Results show that the models predict the personality of unseen utterances better than the mean value baseline, for each Big Five trait. Finally, Chapter 9 presents a novel approach for data-driven language generation with stylistic control, without the computational cost incurred by overgeneration methods. This technique—implemented in the PERSONAGE-PE generator—requires the training of *parameter estimation models* that predict each generation parameter value given target personality scores along the Big Five dimensions. Chapter 9 also presents the first large-scale human evaluation of the linguistic variation produced by a data-driven generator. Results show that the personality conveyed by PERSONAGE-PE's utterances is successfully recognised by human judges, for all Big Five traits apart from conscientiousness. Although PERSONAGE-PE's data-driven models do not improve over the handcrafted parameter settings defined in Chapter 4 when generating extreme personality, parameter estimation models provide an efficient way to (1) generate personality varying over a *continuous* scale and (2) target multiple personality traits *simultaneously*.

Chapter 10 concludes with a discussion of the generalisation of PERSONAGE to new application domains—e.g. interactive narrative systems—while summarising the results, implications and limitations of this research.

Chapter 2

Background

The present work lies at the intersection of two distinct fields of research: personality psychology and natural language processing. As the objective of this chapter is to give the reader background information on both aspects of this thesis, Section 2.1 presents elements of personality psychology underlying the stylistic dimensions that we aim to model, and Section 2.2 reviews findings about how these dimensions affect language. Section 2.3 then details previous work related to user modelling in dialogue, with an emphasis on personality recognition. Finally, Section 2.4 presents an overview of the natural language generation process, as well as recent advances in the control of linguistic variation using rule-based and data-driven techniques.

2.1 Elements of personality psychology

Philosophers and psychologists have tried to identify important dimensions of human behaviour for thousands of years, beginning with Theophrastus in Ancient Greece (B.C. 371-287), who detailed various characters such as 'the Coward', 'the Flatterer' and 'the Unpleasant Man'. In more recent times, the study of personality gained popularity with psychologists such as Sigmund Freud, Carl Jung or Abraham Maslow. Personality has been characterised in many different ways: as a complex interaction between subconscious and conscious processes (psychoanalytic approach, e.g. Freud), as the set of goals of an individual (humanistic approach, e.g. Maslow), as a result of genetic factors (biological approach, e.g. Gray), and as a combination of invariant characteristics (trait theory, e.g. Allport). As mentioned in Section 1.3.2, trait theories have become the most standard approach to personality psychology, we therefore choose to model personality as a set of independent traits. Additionally, we believe that it is the most suitable approach for computational modelling, as it summarises an individual's personality in a concise set of scalar values representing each personality dimension.

2.1.1 The main dimensions of personality

Personality traits describe consistent patterns of individual behaviour [Allport and Odbert, 1936, Norman, 1963]. When talking about a close friend, one can usually come up with hundreds of descriptive adjectives. To be able to reason about personality, psychologists have therefore tried to identify the most *essential* personality traits. The most popular method relies on the *Lexical Hypothesis*, i.e. that any trait important for describing human behaviour has a corresponding lexical token. These lexical tokens are typically adjectives, e.g. *trustworthy, modest, friendly, spontaneous, talkative, dutiful, anxious, impulsive, vulnerable*, etc. Allport and Odbert [1936] collected 17,953 trait terms from English and identified 4,500 as 'stable traits'. This approach led to a great deal of subsequent work over the last century, and a lively debate over the definition of the most essential traits. However, in the last 20 years, a standard framework has emerged of the *Big Five* personality traits [Norman, 1963, Peabody and Goldberg, 1989, Goldberg, 1990], which consist of *extraversion*, *emotional stability*, *agreeableness*, *conscientiousness* and *openness to experience*. Table 2.1 provides a description of each trait. These traits were repetitively obtained empirically by extracting the main components of a factor analysis over adjective descriptors, they are thus considered as the dimensions explaining the most variance of behaviour among people.

Even if they have become a standard in psychology, they suffer from some limitations [Eysenck, 1991, Paunonen and Jackson, 2000]. A first weakness is that the Big Five dimensions are not fully orthogonal, as it was found that extraversion and emotional stability are usually correlated. Furthermore, psychologists argue that many dimensions of personality cannot be represented in this framework, such as honesty, sense of humour, or the level of self-monitoring. For example, self-monitoring individuals tend to adapt their behaviour to what is expected from them. Although they could be represented as a combination of high extraversion

	High	Low
Extraversion	warm, gregarious, assertive, sociable, excitement seeking, active, spontaneous, optimistic, talkative	shy, quiet, reserved, passive, solitary, moody, joyless
Emotional stability	calm, even-tempered, reliable, peaceful, confident	neurotic, anxious, depressed, self-conscious, oversensitive, vulnerable
Agreeableness	trustworthy, friendly, considerate, generous, helpful, altruistic	unfriendly, selfish, suspicious, uncooperative, malicious
Conscientiousness	competent, disciplined, dutiful, achievement striving, deliberate, careful, orderly	disorganised, impulsive, unreliable, careless, forgetful
Openness to experience	creative, intellectual, imaginative, curious, cultured, complex	narrow-minded, conservative, ignorant, simple

Table 2.1: Example adjectives associated with the Big Five traits.

and high agreeableness, it would fail to represent the concept fully. A final criticism is that the Big Five are not based on any underlying theory, but purely extracted from data.

2.1.2 Biological causes

The Big Five framework is a *descriptive model*: it does not provide insight into the underlying mechanisms that cause such traits to be expressed in the population. Nevertheless, there has been some research investigating the biological causes of personality traits. Eysenck et al. [1985] proposed a biologically-motivated model of personality, consisting of three dimensions: extraversion, neuroticism and psychoticism (the PEN model). The first two traits are also part of the Big Five framework, but psychoticism—which is related to tough-mindedness—can be interpreted as a combination of the three remaining traits. Eysenck's theory is based on the *ascending reticular activating system* (ARAS), which is a part of the brain that regulates the level of arousal by controlling the amount of stimulation entering the brain. Eysenck hypothesised the ARAS to be associated with extraversion. For instance, the ARAS of introverts provides them with too much sensory input, they therefore avoid any additional stimulation. On the other hand, extraverts do not receive enough stimuli, hence they look for more sensations to maintain their level of sensory input. Eysenck's hypotheses have been confirmed by his experiments, which show that introverts tend to blink quicker when exposed to light, they turn the volume down sooner when listening to loud music, and they produce more saliva to counteract the acidity of lemon juice [Eysenck et al., 1985]. Eysenck also suggests that introverts divide their short-term memory usage into task-related and self-related concerns. They thus need to access their long-term memory more frequently, which explains why they produce more hesitations and take more time to retrieve lexical items.

Eysenck also hypothesised a biological explanation for neuroticism. This dimension is linked to activation thresholds in the *sympathetic nervous system*, which is the part of the brain responsible for preparing the organism to face danger. Neurotics have a low activation threshold, i.e. they react strongly to a small threat. Their organism is more likely to switch to a 'fight or flight' state, resulting in an increase of the heart rate, sweating and muscular tension. An interesting extreme effect of neuroticism is the panic attack: it occurs when the organism reacts to a threat by amplifying the perception of that threat, creating a chain reaction that paralyses the victim with fear. This can be seen as an unstable feedback loop, like when putting a microphone next to a speaker.

Biological correlates of other Big Five traits have not been studied as extensively. However, DeYoung et al. [2005] have recently found that openness to experience is associated with 'cognitive exploration', which is modulated by dopamine—a hormone that associates novelty with a positive reward, as well as improving memory retrieval capabilities.

Interestingly, each of the Big Five traits has been shown to result from both the environment and genetic factors. Bouchard and McGue [2003] review studies comparing the personality of monozygotic (i.e. identical) twins with dizygotic (i.e. non-identical) twins raised apart, showing that the percentage of the variance accounted for by genetic factors—i.e. the heritability—is approximately 50% for each Big Five trait.

2.2 Language and personality

Why do we believe it might be possible to recognise and convey personality using linguistic cues? Psychologists have documented the existence of such cues by discovering correlations between a range of linguistic variables and personality traits, across a wide range of linguistic levels, including acoustic parameters [Smith et al., 1975, Scherer, 1979], lexical categories [Pennebaker and King, 1999, Pennebaker et al., 2003, Mehl et al., 2006, Fast and Funder, 2007], n-grams [Oberlander and Gill, 2006] and speech-act types [Vogel and Vogel, 1986]. Of all Big Five traits, extraversion has received the most attention from researchers. However, studies focusing systematically on all Big Five traits are becoming more common.

2.2.1 Markers of extraversion

Studies of markers of extraversion generally show that there is a high correlation between extraversion and oral language, especially when the study involves a complex task. Extraverts talk more, louder and more repetitively, with fewer pauses and hesitations. They have higher speech rates, shorter silences, a higher verbal output, a lower type/token ratio and a less formal language, while introverts use a broader vocabulary [Siegman and Pope, 1965, Scherer, 1979, Furnham, 1990, Gill and Oberlander, 2002]. Extraverts also use more positive emotion words, and show more agreements and compliments than introverts [Pennebaker and King, 1999]. Extravert students learning French as a second language produce more back-channels, and have a more implicit style and a lower lexical richness in formal situations. It seems that the more complex the task and the higher the level of anxiety, the easier it is to differentiate between introverts and extraverts [Dewaele and Furnham, 1999].

Heylighen and Dewaele [2002] also note that extraversion is significantly correlated with contextuality, as opposed to formality. Contextuality can be seen as a high reliance on shared knowledge between conversational partners, leading to the use of many deictic expressions such as pronouns, verbs, adverbs and interjections, whereas formal language is less ambiguous and assumes less common knowledge. In order to measure this variation, Heylighen and Dewaele suggest the use of a metric called formality, defined as:

$$F = (\text{noun freq} + \text{adjective freq} + \text{preposition freq} + \text{article freq} - \text{pronoun freq} - \text{verb freq} - \text{adverb freq} - \text{interjection freq} + 100)/2$$

They argue that this measure is the most important dimension of variation between linguistic expressions, as shown in Biber's factor analysis of various genres

[Biber, 1988]. In addition to introversion, the authors also find that formality correlates positively with the level of education and the femininity of the speaker.

Scherer [1979] shows that extraverts are perceived as talking louder and with a more nasal voice, and that American extraverts tend to make fewer pauses, while German extraverts produce more pauses than introverts. Thus personality markers are culture-dependent, even among western societies.

Oberlander and Gill [2006] use content analysis tools and n-gram frequency counts to identify markers in extravert and introvert emails. They replicate previous findings and identify new personality markers such as first person singular pronouns and formal greetings (e.g. *hello*) for introversion, while extraverts use less formal expressions such as *take care* and *hi*.

2.2.2 Markers of other Big Five traits

Pennebaker and King [1999] identify many linguistic features associated with each of the Big Five personality traits. They use the Linguistic Inquiry and Word Count [LIWC; Pennebaker et al., 2001] tool to count word categories of essays written by students whose personality has been assessed using a questionnaire. The authors find small but significant correlations between their linguistic dimensions and personality traits. Neurotics use more first person singular pronouns, more negative emotion words and less positive emotion words. On the other hand, agreeable people express more positive and fewer negative emotions. They also use fewer articles. Conscientious people avoid negations, negative emotion words and words reflecting discrepancies (e.g. *should* and *would*). Finally, openness to experience is characterised by a preference for longer words and words expressing tentativity (e.g. *perhaps* and *maybe*), as well as the avoidance of first person singular pronouns and present tense forms.

Additionally, Mehl et al. [2006] study markers of personality as perceived by observers. They find that the use of words related to insight and the avoidance of past tense indicates openness to experience, and swearing marks disagreeableness. The same authors also show that some linguistic cues vary greatly across gender. For example, males perceived as conscientious produce more filler words, while females do not. Gender differences are also found in markers of self-assessed personality: the use of second person pronouns indicates a conscientious male, but an unconscientious female.

Gill and Oberlander [2003] study correlates of emotional stability in emails: they find that neurotics use more concrete and frequent words. However, they also show that observers do not use those cues correctly, as observer reports of neuroticism correlate negatively with self-reports.

Concerning prosody, Smith et al. [1975] show that speech rate is positively correlated with perceived competence (conscientiousness), and that speech rate has an inverted-U relationship with benevolence (agreeableness), suggesting a need for non-linear computational models.

Some traits have produced more findings than others. A reason for this might be that some are better reflected through language, like extraversion. However, another reason could be that linguistic markers of extraversion are easier to analyse (e.g. verbosity).

The interested reader can find a more extensive review of personality markers of each Big Five trait in Chapter 4, together with a series of hypotheses regarding their ability to control the personality conveyed by a natural language generator.

2.3 User modelling in dialogue

Task-oriented dialogue systems typically try to optimise two aspects of the dialogue's outcome: user satisfaction—e.g. according to questionnaires—and application-dependent task performance metrics, e.g. the number of conversation turns required to successfully book a flight. As users have specific needs and preferences, one way to improve dialogue interaction is to model these individual differences. Previous research has thus investigated various types of *user models*, whose purpose is to categorise each user—either explicitly (e.g. through an elicitation phase) or implicitly (e.g. though keyword spotting)—in order to tailor the interaction to that specific user, or user group. The scenario in Section 1.1, in which the system changes its dialogue strategy based on the user's anxiety, is an example of user modelling. While Chapter 1 provides evidence supporting the utility of personality modelling, most previous work has focused on models of the user's preferences and level of expertise.

2.3.1 Individual preferences

Research on user modelling can be traced back to the GRUNDY system [Rich, 1979], which categorises users in terms of *stereotypes* in order to find and present relevant books to the user. Stereotypes are handcrafted sets of user properties (e.g. the 'educated person' stereotype), which are triggered by specific keywords uttered by the user. Each active stereotype is associated with a confidence score, which can be modified in real-time whenever the system detects a change in user satisfaction. The stereotypes affect user features that are relevant to literary preferences—e.g. the 'educated person' stereotype yields a high value for the 'likes literature' feature—which in turn influence the book selection process. An evaluation shows that users prefer GRUNDY's selection over a randomly selected book.

Walker et al. [2004] model content selection preferences in the MATCH multimodal system, by producing user-tailored restaurant recommendations based on the attributes that are important for the user (e.g. food quality). User preferences for each attribute are modelled as a scalar value collected in an elicitation phase. The MATCH system outputs the restaurant selection maximising the sum of objective attribute ratings (e.g. from a tourist guide) weighted by user preference values, e.g. restaurants with high food quality ratings are only selected if the user expresses a high preference for that attribute. In the same domain, Mairesse and Walker [2005] model the user's syntactic and lexical preferences by training utterance selection models from individual user feedback. Individual models are shown to outperform models trained on the preferences of the average user, suggesting a need for modelling the user's linguistic preferences in dialogue.

2.3.2 Expertise

User expertise has also been modelled in dialogue systems, as users that are familiar with a system interact differently than novice users. Models of the user's expertise can be traced back to the TAILOR question-answering system [Paris, 1988], which provides different content presentation strategies based on the user's domain knowledge. Expertise affects the quantity of information that is presented, as well as the way the information is conveyed, as expert texts typically describe a device in terms of its subparts (i.e. using a *constituency schema*) whereas descriptions aimed at novice readers tend to focus on the device's underlying process.

While TAILOR models the system's response based on the user's knowledge, it assumes that the level of expertise is given to the system. Jokinen and Kanto [2004] propose a speech-based email system in which user expertise is inferred dynamically from dialogue-related variables, such as the number of previous sessions with the user, the number of requests for help, and the number of time-outs. The system can adapt to the user over multiple dialogues, by controlling the level of explicitness of its prompts.

In earlier work, Chu-Carroll [2000] also models user expertise dynamically, resulting in changes in the system's level of initiative. The MIMIC system—which presents movie showtime information—updates the probability of an initiative shift based on user cues, including utterances providing ambiguous or invalid information, or unnecessary repetitions. While MIMIC does not model expertise explicitly, an evaluation shows that its adaptive models improve both user satisfaction and dialogue performance.

2.3.3 Personality

While Reeves and Nass [1996] show that users respond positively to changes in the system's personality (see Chapter 1), they assume that the user's personality is known by the system. As the assessment of the user's personality using questionnaires is an intrusive, time-consuming task, there is a need for automated personality recognition methods. To our knowledge, there are only two other studies on the recognition of personality through language, both focusing on the classification of self-reported personality from text. As these studies are the closest to the one presented in the next chapter, we review them in more detail.

Argamon et al. [2005] learn to classify the personality of students from stream-of-consciousness essays. They associate each essay with a set of linguistic features, consisting of the relative frequencies of function words and word categories based on networks of the theory of systemic functional grammar. They train support vector machine binary classification models predicting the student's level of extraversion and neuroticism, by training their models on the two thirds of the essays with the most extreme personality ratings. A 10-fold cross-validation shows that their models produce 58% correct classifications for both traits. Argamon et al.

identify that relative frequencies of function words are the best predictor for extraversion, suggesting that those that refer to norms are the most salient (e.g. *normally*, *enough*, *particular*). Concerning neuroticism, the feature set characterising appraisal produces by far the best results. Appraisal features are relative frequencies of positive and negative words as well as frequencies of each category in the 'attitude network' (e.g. affect, appreciation, judgement). They find that neurotics tend to use more words related to negative appraisal and affect, but fewer words related to appreciation (e.g. *beautiful*, *innovative*, *hideous*), suggesting that they focus more on their personal feelings.

Oberlander and Nowson [2006] train Naive Bayes and support vector machine classifiers predicting four of the Big Five traits on a corpus of personal weblogs. They follow a bottom-up feature discovery method, by comparing various n-gram feature sets. When building Naive Bayes models using the most frequent bi-grams and tri-grams computed over the full corpus, the authors find that the model of agreeableness is the only one outperforming the baseline (54% accuracy, no level of significance mentioned). When keeping only n-grams that are distinctive of two extreme sets of a given trait, accuracies range from 65% for extraversion to 72% for emotional stability. Finally, when applying an automatic feature selection algorithm to the filtered set, accuracies increase to range from 83% for emotional stability to 93% for agreeableness. In order to be able to compare with Argamon et al., the authors also report experiments where they remove texts with non-extreme personality scores from their corpus, resulting in accuracies up to 100%.

When testing whether these models generalise to a different corpus of weblogs, Nowson and Oberlander [2007] report binary classification accuracies ranging from 55% for extraversion to 65% for conscientiousness. Interestingly, models trained on the most extreme instances of the original corpus outperform models trained on the full corpus, although no level of significance is mentioned. These studies show that n-grams can be appropriate to model self-reports of personality.

2.4 Modelling individual differences in natural language generation

While most commercial dialogue systems still use pre-recorded prompts, the past 20 years have seen the emergence of natural language generation (NLG) components, that aim at automatically converting the system's communicative goal into an utterance. As dialogue systems become more flexible, pre-recording all the possible outputs becomes impractical, and language generators become necessary to produce a wide range of utterances.

Most NLG systems have focused on the production of grammatical and natural language, usually in the written form [Bateman, 1995, Elhadad and Robin, 1996, Reiter and Dale, 2000, Langkilde-Geary, 2002]. The rules for the production of spoken language are not as well-defined, as spoken language is more dependent on the speaker's linguistic style, e.g. level of formality, dialect or sociolect (see Section 1.3.2). Furthermore, natural dialogue involves a large range of pragmatic effects that are not found in written text, e.g. grounding, conversational implicature, or social hedging expressions. While most NLG systems ignore linguistic variation by producing a single output satisfying the system's communicative goal, a recent line of research investigates techniques for controlling the generated linguistic variation—referred to as *stylistic generation*—which is also the focus of the second part of this thesis (see Chapters 4 to 9).

2.4.1 Early work: ELIZA and PARRY

The ELIZA chatterbot is usually considered as the first interactive dialogue system [Weizenbaum, 1966]. ELIZA imitates a psychotherapist's initial interview, by systematically asking the user to follow up on his or her input. For example, if the user says *'I have a headache'*, ELIZA would reply *'Why do you say you have a headache?'* or *'Please elaborate on that'*. This particular domain allows the production of long dialogues without requiring specific knowledge sources. Although the language generation mechanism is limited to canned phrases, Weizenbaum [1976] found that users enjoy interacting with ELIZA, and they even attribute emotions to the system.

While ELIZA's language is generally agreeable and supportive, the PARRY chatterbot uses a more disagreeable, anxious tone, simulating a paranoid schizophrenic [Colby, 1975], e.g. to the question *'Why are you in the hospital?'* it would answer *'I shouldn't be here'*. The PARRY system involves more advanced natural language processing techniques, as it models the beliefs and intentions of a paranoiac. Its output is thus a function of the user's input, the topic under discussion, and its relation to PARRY's set of paranoid beliefs.

2.4.2 The standard NLG architecture

While ELIZA and PARRY were successful at conveying specific personalities, their generation capabilities were limited to canned text. Since then, many candidate NLG architectures have emerged. Researchers have tried to identify commonalities between them, which has resulted in a 'consensus' NLG architecture [Reiter and Dale, 2000]. Figure 2.1 illustrates the pipelined architecture, which consists of a sequence of components that successively transform the input meaning representation into an utterance string:

- **Content planning:** The content planner refines the input communicative goals, by selecting and structuring the propositional content, e.g. using a rhetorical structure tree [Mann and Thompson, 1988]. Its output is a content plan, which defines *what* will be talked about in the utterance. The content planner performs two main tasks:
  - **Content selection:** The choice of the propositional elements to be expressed in the utterance;
  - **Rhetorical structuring:** The definition of the relations between each propositional element, e.g. using a CONTRAST rhetorical relation for contrasting two propositions.
- **Sentence planning:** The sentence planner—also known as the *micro-planner*—is responsible for converting the content plan into a syntactic representation, which typically includes the following tasks:
  - **Content ordering:** The choice of linear ordering of the propositional content to be expressed;

  - **Syntactic template selection:** The choice of what syntactic template to select for expressing each proposition, typically chosen from a hand-crafted generation dictionary;
  - **Aggregation:** The combination of the propositions' syntactic representations to form the utterance's full syntactic structure, e.g. by selecting specific cue words for expressing the content plan's rhetorical relations;
  - **Referring expression generation**: The substitution of named entities for referring expressions, e.g. pronouns or generic nouns;
  - **Lexical choice:** The selection of the final lexical item for each content word in the utterance.

- **Surface realisation:** The transformation of the utterance's syntactic representation into a surface string, e.g. by inserting function words and applying morphological rules.

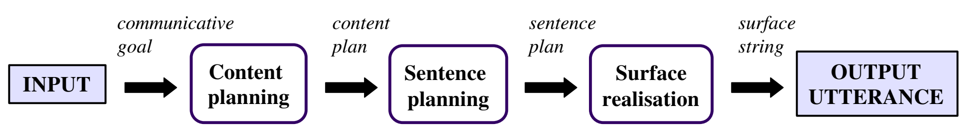


Figure 2.1: High-level pipelined architecture of a natural language generator.

There are some limitations to this architecture, such as the impossibility for a given component to inform any preceding component in the pipeline. This makes it impossible for the surface realiser to affect the content selection process for example, which would be required to satisfy global constraints on the utterance's surface string, such as a maximum length. Nevertheless, similar architectures have been widely used in rule-based generation systems. Therefore, we use this architecture as a building block for generating language conveying personality in the second part of this thesis (see Chapter 4).

2.4.3 Template and rule-based stylistic generation

We now turn to an overview of previous research on the generation of linguistic variation. This section focuses on systems that control the output style using hand-crafted templates as well as more complex rule-based techniques.

2.4.3.1 Pragmatic effects

Research on pragmatic effects in natural language generation is usually traced back to the PAULINE system [Hovy, 1988]. PAULINE's input is a set of pragmatic goals, such as the opinion or style to be conveyed. To reduce the complexity of the rules involved, linguistic features are associated with a small set of intermediary *rhetorical goals*, which are combined to generate various pragmatic effects. For example, the rhetorical goals of low formality, high force and high partiality produce a 'no-nonsense' effect. Other pragmatic effects include the control of the topic's subjective connotation, the confusion induced in the hearer, or the distance between the hearer and the speaker. Rhetorical goals affect many levels of the NLG pipeline, including content selection and grouping, template selection, clause inclusion and organisation, and lexical choice.

While PAULINE is the first NLG system with explicit pragmatic modelling, some of its limitations are worth looking at, as they have not been entirely addressed yet. First, the rules associating the rhetorical goals with linguistic markers are based on findings from the linguistics literature as well as intuition. With the current level of technology, it might be possible to acquire such rules from data, in order to increase the system's coverage. A second limitation is that each goal is represented as a discrete variable, while the intensity of each goal could be modelled continuously, e.g. using appropriate weighting. Finally, PAULINE's pragmatic effects have not been evaluated, it is thus not clear whether they would be perceived correctly by humans.

More recently, Fleischman and Hovy [2002] present a system that controls the subjectivity of its output, by selecting the utterance minimising a distance measure between the connotation of individual phrases and a scalar vector representing the attitude of the user towards the phrase's object. For example, the system would choose the template utterance *'Y smashed into X'* over *'X was hit by Y's car'* if the user is known to dislike Y. An evaluation shows that human perceptions of the attitude of the speaker towards the utterance's object correlate significantly with the system's emotional target ($r = .66$).

2.4.3.2 Linguistic style

DiMarco and Hirst [1993] model linguistic style using a set of *stylistic grammars*. These grammars associate linguistic markers with primitive stylistic elements, which in turn are mapped to stylistic goals over three dimensions: clarity/obscurity, concreteness/abstraction, and dynamism/staticness. For example, a 'heteropoise' sentence—i.e. using multiple grammatical forms—results in a low clarity but a high concreteness.

Green and DiMarco [1993] use stylistic grammars to influence the language generation process. Following a top-down approach, a stylistic control module applies the grammar to produce a list of primitive stylistic elements based on the target stylistic goals (e.g. use of conjunctive clause for concreteness). This list is then consulted by the generator whenever it has to make a decision regarding a stylistic element (e.g. clause aggregation).

While the stylistic control is more grounded into formal linguistics than in PAULINE, it does suffer from the same limitations: it requires a large amount of handcrafted rules, and it lacks any human evaluation.

2.4.3.3 Politeness

The most influential model of politeness is Brown and Levinson's politeness theory [1987]. The authors model politeness as a mechanism for maintaining the hearer's *face*, i.e. the hearer's freedom of action (*negative face*, or desire for autonomy) and the hearer's belief that his or her wants are seen as desirable (*positive face*, or desire for approval). For example, ordering the hearer to pass the salt threatens the hearer's negative face, while ignoring the hearer's point of view affects his or her positive face. Brown and Levinson model the choice of politeness strategy based on the seriousness of the face-threatening act, which is dependent on the social distance between the hearer and the speaker, the power of the speaker over the hearer, as well as the type of speech act being uttered.

Walker et al. [1997] propose an implementation of this theory in a computational model. Their system evaluates the threat to the user's face during the dialogue, and selects the output utterance according to the current politeness strategy. While this work represents the first computational modelling of an existing politeness theory, the language generation mechanism is limited to the selection of syntactic forms from a generation dictionary, and there is no evaluation of the effect of the politeness strategy on the user.

Wilkie et al. [2005] present such an evaluation in their telephone banking system. The system can choose between three politeness strategies to interrupt the user with offers: negative face redressment (e.g. *'I'm very sorry to interrupt...'*), positive face redressment (e.g. *'I know you won't mind me cutting in...'*), as well as a bald interruption (no politeness form). The negative face-redressive strategy was perceived as more formal, more apologetic and more polite. The positive face-redressive strategy was seen as the most manipulative of the three and it was considered significantly more intrusive and patronising than the bald strategy. Although the authors find that 66% of the users perceive the negative face-redressive strategy as the most polite, 54% still prefer the bald strategy, as it was described as less patronising, less intrusive, less formal, more honest and more professional. Out of the remaining users, 29% preferred the negative face-redressive strategy, mainly because it was seen as polite and apologetic. While those results show that most users favour efficiency over politeness, almost a third of the users prefer interacting with a polite system, suggesting that user modelling would improve overall user satisfaction.

Porayska-Pomsta and Mellish [2004] apply Brown and Levinson's theory to the generation of corrective feedback in an intelligent tutoring system, by re-interpreting the notion of face in terms of the student's need for autonomy (by controlling the feedback's level of guidance) and approval (by controlling the feedback's positivity). Their system takes a series of situational factors as an input (e.g. the remaining time, and estimates of the student's ability and confidence) which are converted to a scalar vector representing the desired level of autonomy and approval of the answer, using a Bayesian network derived from teachers' feedback. The utterance with the closest scalar values is then selected from a hand-labelled generation dictionary based on sample human-human tutorial dialogues. An evaluation shows that human tutors rate the system's output as significantly more suitable than the system's less preferred answer, while the system's output is not perceived as significantly different from a handcrafted human answer.

Wang et al. [2005] also study the application of Brown and Levinson's theory in a tutoring system, however they focus on the influence of politeness on the learning outcome. They conduct a Wizard-of-Oz experiment in which the system presents instructions for using a factory modelling tool, which are either negative or positive face-redressive (e.g. *'Why don't we save our factory now?'*) or direct (e.g. *'Save the factory now'*). Interestingly, results show that participants exposed to the polite system perform significantly better according to a learning outcome questionnaire, although they do not report any preference for the polite system.

While politeness models have yet to be implemented in a flexible language generator, previous research using handcrafted templates shows that politeness modelling can affect both the user's perception of the system and the dialogue outcome.

2.4.3.4 Personality and embodied conversational agents

Previous work on personality generation has typically been associated with embodied conversational agents (ECA).

Ball and Breese [1998] build a Bayesian network to model the effect of the agent's personality (i.e. dominance and friendliness) and emotions (i.e. valence and arousal) on its behaviour. The personality values affect a layer of variables determining the paraphrase to be selected by the system, such as the language's strength, positivity and terseness. André et al. [2000] model the personality of multiple conversational agents based on the extraversion, agreeableness and openness to experience dimensions of the Big Five framework, assuming they are the most important for social interaction. Again, templates are annotated with intermediary variables (e.g. force) which in turn are associated with the personality traits (e.g. extravert agents use more forceful language, and they show more initiative in dialogue). The NECA system is a multimodal language generator that models pragmatic effects and personality [Piwek, 2003]. Information about the character's personality is passed from one module to the other in order to produce consistent behaviour across modes (e.g. language, speech and gesture), while the way personality affects language is encoded in a generation grammar.

Cassell and Bickmore [2003] extend their REA real estate agent with smalltalk generation capabilities, which is hypothesised to increase the user's trust in the system. Interestingly, they observe large perceptual variations between user groups with different personalities. Extravert users feel that they know REA better if she produces social language, resulting in a more satisfying interaction. On the other hand, introvert users are much less affected by REA's smalltalk, which even results in slightly lower satisfaction ratings.

There has been a lot of work on personality modelling in ECA's, however the natural language generation task is typically avoided by using hand-labelled templates. Personality is generally modelled at the discourse level rather than lower in the NLG pipeline. Additionally, personality is usually modelled using heuristic rules, rather than findings from psychological studies. Finally, we do not know of any study evaluating whether the personality of generated utterances is recognised by human users.

2.4.4 Data-driven stylistic generation

The systems presented in the previous section rely on handcrafted templates or rules, which are typically limited to one domain and hard to maintain. Furthermore, linguistic variation is manipulated using linguistic markers that are mostly based on the author's intuition rather than scientific experimentation. Recent work in NLG attempts to address these issues by using large collections of text to inform the generation process. As this data is generally used to train a statistical model predicting characteristics of the output utterance, this paradigm is referred to as *trainable* or *data-driven generation*.

2.4.4.1 Overgenerate and select methods

Early work in data-driven generation has focused on utterance quality, such as the NITROGEN generator and its successor HALOGEN [Langkilde and Knight, 1998, Langkilde-Geary, 2002]. HALOGEN is based on the *overgenerate and select* generation method: many candidate utterances are generated using a large set of rules, and the system selects the utterance with the highest probability according to an n-gram language model trained on newspaper articles. The base generator is allowed to produce ungrammatical candidates, as the language model is responsible for ensuring grammaticality and naturalness. While HALOGEN was trained on a large corpus of Wall Street Journal articles (250 million words), it is not clear whether it can be applied to spoken language generation, as spoken language corpora are typically much smaller. The main advantage of HALOGEN's approach is its flexibility: the language model can be modified independently of the base generator, e.g. in order to change the style of the output.

Inkpen and Hirst [2004] exploit this flexibility by adding a *near-synonym choice module* to HALOGEN, which weights the n-gram probability estimates using a similarity measure between input stylistic preferences and the characteristics of the utterance's content words. These characteristics are extracted from a dictionary; they consist of information about the word's formality, connotation, and how it affects the phrase's agent. This work represents a first hybrid method that uses data-driven generation to optimise the utterance's naturalness and a dictionary to control its lexical style. However, a limitation is that the variation is modelled on a word-by-word basis, which might produce more inconsistencies than systems based on deeper primitive elements such as syntactic frames.

Isard et al. [2006] use the same methodology as HALOGEN to model personality and alignment in dialogue between conversational agents. Their CRAG-2 system is based on a linear combination of n-gram models, including a general-domain model trained on conversations from the Switchboard corpus, and models trained on a corpus of weblogs labelled with the author's personality. The system models linguistic alignment using a 'cache' language model based on the conversational partner's previous utterance. Whereas Inkpen and Hirst's technique uses an external knowledge source, CRAG-2's selection phase is purely data-driven. While CRAG-2 has yet to be evaluated, a potential limitation is that n-gram models tend to produce an 'average' style by filtering out infrequent linguistic cues, it is therefore not clear whether they can successfully model speakers with extreme, infrequent personality traits.

While previous generators are trained on corpora, the SPOT sentence planner and its successor SPARKY are trained on user ratings of output utterances [Walker et al., 2002, Stent et al., 2004]. While n-grams model small word sequences from the surface realisation, SPOT's statistical ranker predicts the utterance quality from deeper content and syntactic features. Although the main limitation of this approach is that it requires the collection of ratings for hundreds of output utterances, the resulting data does not contain the level of noise found in general-domain corpora. The authors show that user ratings of the utterance selected by SPoT are only 5% worse than the highest rated candidate utterance, while a large-scale human evaluation shows that SPoT performs as well as an existing handcrafted generator in the same domain.

The interested reader can find an implementation of the overgenerate and select method to convey personality in dialogue in Chapter 8, which compares models trained on user feedback with models trained on general-domain corpora.

2.4.4.2 Direct control of the generation process

Paiva and Evans [2005] present a technique for controlling the output linguistic variation that does not overgenerate.[1] Stylistic dimensions are first extracted from a corpus by doing a factor analysis of general linguistic features. The second step is for the generator to randomly generate many utterances, while recording the corresponding generation decisions as well as the utterance's score along each stylistic dimension. This data is then used to train a linear regression model to predict the stylistic scores from the generation decisions. At generation time, the generator solves the system of linear equations at each choice point to find the generation decision yielding the closest stylistic scores to the target style. While there is no guarantee that the extracted stylistic dimensions can be easily interpreted by humans, this method can be adapted to other variation dimensions as long as they can be estimated for each training instance.

Belz [2005b] compares various data-driven techniques for generating weather forecasts. The generation process is modelled as a context-free grammar, which is used to parse a corpus of weather forecasts. Frequency counts of each derivation are then used to estimate the probability of each generation decision. Results show that a greedy generation policy—i.e. selecting the most likely decision at every choice point—performs as well as HALOGEN's n-gram ranking technique, but at a much lower computational cost. Belz compares generation grammars trained on corpora of individual forecasters [Belz, 2005a], showing that n-gram ranking models overfit more than models based on generation decisions. While this suggests that this technique is suitable for modelling individual variation, a limitation

[1] It only requires an overgeneration phase during the system's development, not at generation time.

is that most rule-based generators implementing the standard NLG architecture detailed in Section 2.4.2 cannot be reduced to a context-free grammar, which makes the automated annotation of generation decisions difficult.

Chapter 9 evaluates a novel data-driven approach to personality generation that extends the method presented by Paiva and Evans [2005], which also estimates the optimal generation decisions statistically at generation time.

2.5 Summary

This chapter has presented an overview of three distinct fields: personality psychology, user modelling and natural language generation. The Big Five personality traits were shown to be important variation dimensions of human behaviour, including language production. The rest of this thesis will detail how personality can be used to (1) model dialogue system users using linguistic cues (Chapter 3) and (2) control the user's perception of the system (Chapters 4 to 9).

While previous work on user modelling has mainly focused on models of user preferences and expertise, we believe that personality dimensions provide a useful framework for tailoring the interaction to specific groups of users. The next chapter thus investigates data-driven techniques for automated personality recognition, by providing the first evaluation of a personality recogniser trained on spoken language data and predicting personality over a continuous scale.

Natural language generation research has started to focus on methods for modelling pragmatic effects, such as politeness or linguistic style. We propose to control the generation process using the same parameters as those used to model the user—i.e. personality dimensions—thus making it possible to test the effect of user adaptation. The second part of the thesis presents various methods for conveying personality traits through language, using both rule-based and data-driven techniques to extend on previous work [Paiva and Evans, 2005, Isard et al., 2006], as well as the first human evaluation of a data-driven stylistic generation method.

Part I

Recognising the User's Personality in Dialogue

Chapter 3

Personality Recognition from Linguistic Cues

Chapter 1 provides evidence for the utility of personality modelling in dialogue systems, as personality affects many aspects of interpersonal behaviour, such as leadership ability, attitude towards machines, general job performance and teacher effectiveness (see Section 1.4.1). Furthermore, the user's personality was shown to greatly affect the perception of the system [Reeves and Nass, 1996, Cassell and Bickmore, 2003], suggesting the need for controlling the personality of the system based on the personality of the user.

3.1 Adapting to the user's personality

Previous work on alignment in dialogue has focused on lexical items and syntax [Brennan, 1996, Branigan et al., 2000], while suggesting that alignment occurs at all levels of language production [Pickering and Garrod, 2004]. However, evidence presented in Section 1.4 suggests that it is worth exploring alignment at the personality level, which also occurs between humans [Byrne and Nelson, 1965]. We therefore make the hypothesis that alignment is usefully mediated by a concise set of psychologically-motivated personality dimensions: the *Big Five personality traits*. As detailed in Section 2.1.1, the Big Five framework models personality using five dimensions:

- Extraversion vs. introversion (sociable, assertive, playful vs. reserved, shy)
- Emotional stability vs. neuroticism (calm, unemotional vs. insecure, anxious)
- Agreeableness (friendly, cooperative vs. antagonistic, faultfinding)
- Conscientiousness (self-disciplined, organised vs. inefficient, careless)
- Openness to experience (intellectual, insightful vs. shallow, unimaginative)

Figure 3.1 illustrates how personality modelling can be incorporated within a dialogue system architecture. As part of the input understanding phase, a *personality recogniser* is responsible for inferring (or updating) the user's personality based on linguistic and prosodic cues. This task is the focus of the rest of this chapter. A *personality adaptation* module then selects the personality to be conveyed by the system, based on the user's personality and the dialogue task, e.g. an introvert user is likely to benefit from an extravert tutoring system. The model of the system's personality can then inform all aspects of the system's output, such as dialogue strategy selection, natural language generation and speech synthesis. Techniques for conveying the system's target personality through language are discussed in the second part of this thesis (see Chapter 4).

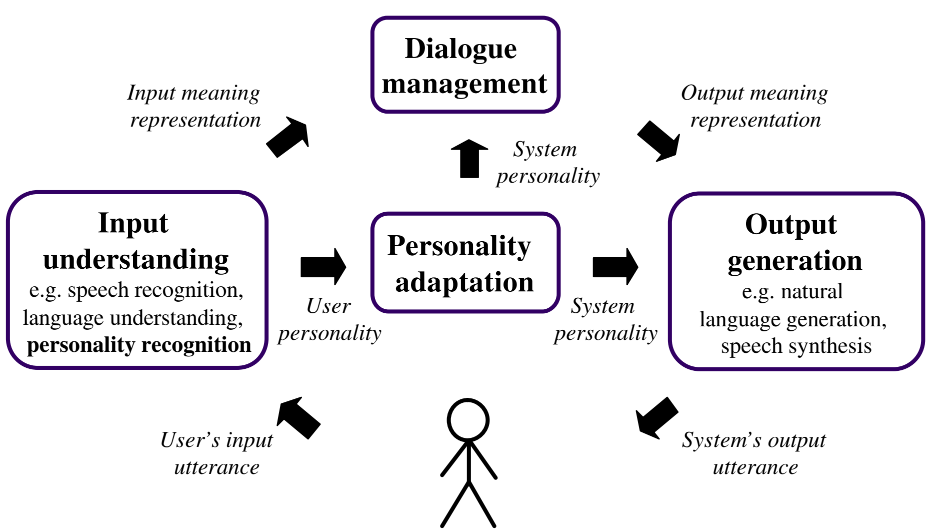


Figure 3.1: High-level architecture of a dialogue system with personality modelling.

While in some applications it would be possible to acquire personality information by asking the user directly [John et al., 1991, Costa and McCrae, 1992], here we explore whether it is possible to acquire personality models for the Big Five personality traits by observation of individual linguistic outputs in text and conversation. We thus make the hypothesis that the personality of an individual can be learnt from a small number of utterances. While this hypothesis will need to be evaluated in future work, new applications are likely to provide a large amount of individual user inputs—e.g. intelligent tutoring systems or home automation controllers—hence allowing for more accurate modelling.

Section 3.2 overviews the methods we use to automatically train personality recognition models, using both conversation and written language samples, and both self-reports and observer ratings of personality traits. We explore the use of classification models (Section 3.3), regression models (Section 3.4) and ranking models (Section 3.5), as well as the effect of different feature sets on model accuracy. The results show that for some traits, any type of statistical model performs significantly better than the baseline, but ranking models perform best overall. In addition, models trained on observer personality scores perform better than models trained using self-reports, and the optimal feature set is dependent on the personality trait. The rules learnt by the models confirm previous findings linking language and personality, while revealing many new linguistic markers.

3.2 Experimental method

We conduct a set of experiments to examine whether automatically trained models can be used to recognise the personality of unseen subjects. Our approach can be summarised in five steps:

1. Collect individual corpora;
2. Collect associated personality ratings for each participant;
3. Extract relevant features from the corpora;
4. Build statistical models of the personality ratings based on the features;
5. Test the learnt models on the linguistic outputs of unseen individuals.

The following sections describe each of these steps in more detail.

3.2.1 Sources of language and personality

Learning to recognise the user's personality in dialogue seems to require in-domain data, i.e. transcripts of conversations with the system. However, the collection of a large amount of personality-annotated dialogues is an expensive task, and the resulting models are likely to overfit and generalise poorly to other domains. The present work focuses on a more general approach, by making the hypothesis that out-of-domain data—such as writings, conversation transcripts, emails or weblogs—can be used to build general, re-usable personality recognition models. In order to test this hypothesis, this chapter focuses on two distinct sources of language: personal writings (stream-of-consciousness essays) and daily-life conversations.

Introvert	**Extravert**
I've been waking up on time so far. What has it been, 5 days? Dear me, I'll never keep it up, being such not a morning person and all. But maybe I'll adjust, or not. I want internet access in my room, I don't have it yet, but I will on Wed??? I think. But that ain't soon enough, cause I got calculus homework [...]	I have some really random thoughts. I want the best things out of life. But I fear that I want too much! What if I fall flat on my face and don't amount to anything. But I feel like I was born to do BIG things on this earth. But who knows... There is this Persian party today.
Neurotic	**Emotionally stable**
One of my friends just barged in, and I jumped in my seat. This is crazy. I should tell him not to do that again. I'm not that fastidious actually. But certain things annoy me. The things that would annoy me would actually annoy any normal human being, so I know I'm not a freak.	I should excel in this sport because I know how to push my body harder than anyone I know, no matter what the test I always push my body harder than everyone else. I want to be the best no matter what the sport or event. I should also be good at this because I love to ride my bike.

Table 3.1: Extracts from the essays corpus, for participants rated as extremely introvert, extravert, neurotic and emotionally stable.

The first corpus contains 2,479 'stream-of-consciousness' essays from psychology students (1.9 million words), who were told to write whatever comes into their mind for 20 minutes. The data was collected and analysed by Pennebaker and King [1999]; a sample is shown in Table 3.1. Personality was assessed by asking each student to fill in the Big Five Inventory questionnaire [John et al., 1991], which asks participants to evaluate on a 5 point scale how well their personality matches a series of descriptions.

Introvert	Extravert
- Yeah you would do kilograms. Yeah I see what you're saying. - On Tuesday I have class. I don't know. - I don't know. A16. Yeah, that's kind of cool. - I don't know. I just can't wait to be with you and not have to do this every night, you know? - Yeah. You don't know. Is there a bed in there? Well ok just...	- That's my first yogurt experience here. Really watery. Why? - Damn. New game. - Oh. - That's so rude. That. - Yeah, but he, they like each other. He likes her. - They are going to end up breaking up and he's going to be like.
Unconscientious	**Conscientious**
- With the Chinese. Get it together. - I tried to yell at you through the window. Oh. xxxx's fucking a dumb ass. Look at him. Look at him, dude. Look at him. I wish we had a camera. He's fucking brushing his t-shirt with a tooth brush. Get a kick of it. Don't steal nothing.	- I don't, I don't know for a fact but I would imagine that historically women who have entered prostitution have done so, not everyone, but for the majority out of extreme desperation and I think. I don't know, i think people understand that desperation and they don't see [...]

Table 3.2: Extracts from the EAR corpus, for participants rated as extremely introvert, extravert, unconscientious and conscientious. Only the participants' utterances are shown.

The second source of data consists of conversation extracts recorded using an Electronically Activated Recorder (EAR), collected by Mehl et al. [2006]. To preserve the participants' privacy, only random snippets of conversation were recorded. This corpus is much smaller than the essays corpus (96 participants for a total of 97,468 words and 15,269 utterances). While the essays corpus consists only of texts, the EAR corpus contains both sound extracts and transcripts. This corpus therefore allows us to build models of personality recognition from speech. Only the *participants'* utterances were transcribed (not those of their conversational partners), making it impossible to reconstruct whole conversations. Nevertheless, the conversation extracts are less formal than the essays, and personality may be best observed in the absence of behavioural constraints. Table 3.3 shows that while the essays corpus is much larger than the EAR corpus, the amount of data per subject is comparable, i.e. 766 words per subject for the essays and 1,015 for the EAR corpus. Table 3.2 shows examples of conversations from the EAR corpus for different personality traits.

For personality ratings, the EAR corpus contains both self-reports and ratings from 18 independent observers. Psychologists use self-reports to facilitate evaluat-

Dataset	Essays	EAR
Source of language	Written	Spoken
Personality reports	Self reports	Self and observer
Number of words	1.9 million	97,468
Subjects	2,479	96
Words per subject	766.4	1,015.3

Table 3.3: Comparison of the essays and EAR corpora.

ing the personality of a large number of participants, and there are a large number of standard self-report tests. Observers were asked to make their judgements by rating descriptions of the Big Five Inventory [John and Srivastava, 1999] on a 7 point scale (from *strongly disagree* to *strongly agree*), without knowing the participants. Observers were divided into three groups, each rating one third of the participants, after listening to each participant's entire set of sound files (130 files on average). The personality assessment was based on the audio recordings, which contain more information than the transcripts (e.g. ambient sounds, including captured conversations). Mehl et al. [2006] report strong inter-observer reliabilities across all Big Five dimensions (intraclass correlations based on one-way random effect models: mean $r = 0.84$, $p < .01$). The observers' ratings were averaged for each participant, to produce the final scores used in our experiments.

Interestingly, the average correlations between frequency counts from psycholinguistic word categories and the Big Five personality dimensions were considerably larger in the EAR corpus than with the student essays studied by Pennebaker and King. Moreover, the correlations reported by Mehl et al. seem to be higher for observer reports than for self-reports. Based on this observation, we hypothesise that models of observed personality will outperform models of self-assessed personality.

3.2.2 Features

The features used in the experiments are motivated by previous psychological findings about correlations between measurable linguistic factors and personality traits. Features are divided into subsets depending on their source and described in the subsections below. The total feature set is summarised in Table 3.5. The experimental results given in Sections 3.3, 3.4 and 3.5 examine the effect of each feature subset on model accuracy.

3.2.2.1 Content and syntax

Feature	Type	Example
Anger words	LIWC	*hate, kill, pissed*
Metaphysical issues	LIWC	*God, heaven, coffin*
Physical states/functions	LIWC	*ache, breast, sleep*
Inclusive words	LIWC	*with, and, include*
Social processes	LIWC	*talk, us, friend*
Family members	LIWC	*mom, brother, cousin*
Past tense verbs	LIWC	*walked, were, had*
References to friends	LIWC	*pal, buddy, coworker*
Imagery of words	MRC	Low: *future, peace* - High: *table, car*
Syllables per word	MRC	Low: *a* - High: *uncompromisingly*
Concreteness	MRC	Low: *patience, candor* - High: *ship*
Frequency of use	MRC	Low: *duly, nudity* - High: *he, the*

Table 3.4: Examples of LIWC word categories and MRC psycholinguistic features [Pennebaker et al., 2001, Coltheart, 1981]. MRC features associate each word with a numerical value.

We extracted a set of linguistic features from each essay and conversation transcript, starting with frequency counts of 88 word categories from the Linguistic Inquiry and Word Count (LIWC) utility [Pennebaker et al., 2001]. These features include both syntactic (e.g. ratio of pronouns) and semantic information (e.g. positive emotion words), which were validated by expert judges. Some LIWC features are illustrated in Table 3.4. Pennebaker and King [1999] previously found significant correlations between these features and each of the Big Five personality traits. Relevant word categories for extraversion include social words, emotion words, first person pronouns, and present tense verbs. Mehl et al. [2006] showed that LIWC features extracted from the EAR corpus were significantly correlated with both self and observer reports of personality. While these correlational studies suggest that it is possible to detect personality from linguistic cues, they do not take prosodic information into account. This chapter aims at showing whether statistical models trained on this data can predict the personality of *unseen* subjects, using both lexical and prosodic information.

We also added 14 additional features from the MRC psycholinguistic database [Coltheart, 1981], which contains statistics for over 150,000 words, such as estimates of the age of acquisition, frequency of use and familiarity. The MRC feature set was previously used by Gill and Oberlander [2002], who showed that extraversion is negatively correlated with concreteness. Concreteness also indicates neu-

LIWC FEATURES [Pennebaker et al., 2001]:

· **Standard counts:**
- Word count (WC), words per sentence (WPS), type/token ratio (Unique), words captured (Dic), words longer than 6 letters (Sixltr), negations (Negate), assents (Assent), articles (Article), prepositions (Preps), numbers (Number)
- Pronouns (Pronoun): 1^{st} person singular (I), 1^{st} person plural (We), total 1^{st} person (Self), total 2^{nd} person (You), total 3rd person (Other)

· **Psychological processes:**
- Affective or emotional processes (Affect): positive emotions (Posemo), positive feelings (Posfeel), optimism and energy (Optim), negative emotions (Negemo), anxiety or fear (Anx), anger (Anger), sadness (Sad)
- Cognitive Processes (Cogmech): causation (Cause), insight (Insight), discrepancy (Discrep), inhibition (Inhib), tentative (Tentat), certainty (Certain)
- Sensory and perceptual processes (Senses): seeing (See), hearing (Hear), feeling (Feel)
- Social processes (Social): communication (Comm), other references to people (Othref), friends (Friends), family (Family), humans (Humans)

· **Relativity:**
- Time (Time), past tense verb (Past), present tense verb (Present), future tense verb (Future)
- Space (Space): up (Up), down (Down), inclusive (Incl), exclusive (Excl)
- Motion (Motion)

· **Personal concerns:**
- Occupation (Occup): school (School), work and job (Job), achievement (Achieve)
- Leisure activity (Leisure): home (Home), sports (Sports), television and movies (TV), music (Music)
- Money and financial issues (Money)
- Metaphysical issues (Metaph): religion (Relig), death (Death), physical states and functions (Physcal), body states and symptoms (Body), sexuality (Sexual), eating and drinking (Eating), sleeping (Sleep), grooming (Groom)

· **Other dimensions:**
- Punctuation (Allpct): period (Period), comma (Comma), colon (Colon), semi-colon (Semic), question (Qmark), exclamation (Exclam), dash (Dash), quote (Quote), apostrophe (Apostro), parenthesis (Parenth), other (Otherp)
- Swear words (Swear), nonfluencies (Nonfl), fillers (Fillers)

MRC FEATURES [Coltheart, 1981]:

Number of letters (Nlet), phonemes (Nphon), syllables (Nsyl), Kucera-Francis written frequency (K-F-freq), Kucera-Francis number of categories (K-F-ncats), Kucera-Francis number of samples (K-F-nsamp), Thorndike-Lorge written frequency (T-L-freq), Brown verbal frequency (Brown-freq), familiarity rating (Fam), concreteness rating (Conc), imagery rating (Imag), meaningfulness Colorado Norms (Meanc), meaningfulness Paivio Norms (Meanp), age of acquisition (AOA)

UTTERANCE TYPE FEATURES:

Ratio of commands (Command), prompts or back-channels (Prompt), questions (Question), assertions (Assertion)

PROSODIC FEATURES:

Average, minimum, maximum and standard deviation of the voice's pitch in Hz (Pitch-mean, Pitch-min, Pitch-max, Pitch-stddev) and intensity in dB (Int-mean, Int-min, Int-max, Int-stddev), voiced time (Voiced) and speech rate (Word-per-sec)

Table 3.5: Description of all features, with feature labels in brackets.

roticism, as well as the use of more frequent words [Gill and Oberlander, 2003]. Table 3.4 shows examples of MRC scales. Each MRC feature is computed by averaging the feature value of all the words in the essay or conversation extract. Part-of-speech tags are computed using MINIPAR [Lin, 1998] to identify the correct entry in the database among a set of homonyms.

3.2.2.2 Utterance type

Various facets of personality traits seem to depend on the level of initiative of the speaker and the type of utterance used (e.g. assertiveness, argumentativeness, inquisitiveness, etc.). For example, extraverts are more assertive in their emails [Gill and Oberlander, 2002], while extravert second language learners were shown to produce more back-channel behaviour [Vogel and Vogel, 1986]. We therefore introduced features characterising the types of utterance produced. We automatically tagged each utterance of the EAR corpus with speech act categories suggested by Walker and Whittaker [1990], using heuristic rules based on each utterance's parse tree:

- Command: utterance using the imperative form, a command verb (e.g. *must*, *have to*) or a yes/no second person question with a modal auxiliary like *can*;
- Prompt: single word utterance used for back-channelling (e.g. *yeah*, *ok*, *huh*, etc.);
- Question: interrogative utterance which is not a command;
- Assertion: any other utterance.

We evaluated the automatic tagger by applying it to a set of 100 hand-labelled utterances randomly selected in the EAR corpus. We obtain 88% of correct labels, which are mostly assertions. Table 3.6 summarises the partition and the evaluation results (recall) for each utterance type. For each utterance type, the corresponding feature value is the ratio of the number of occurrences of that utterance type to the total number of utterances in each text.

Label	**Fraction**	**Labelling recall**
Assertion	73.0%	0.95
Command	4.3%	0.50
Prompt	7.0%	0.57
Question	15.7%	1.00
All	100%	0.88

Table 3.6: Partition of the utterance types automatically extracted from the EAR corpus, and classification accuracies (recall) on a sample of 100 hand-labelled utterances.

3.2.2.3 Prosody

Chapter 2 has shown that personality also affects speech production (Section 2.2). Extraversion is associated with more variation of the fundamental frequency [Scherer, 1979], with a higher voice quality and intensity [Mallory and Miller, 1958], and with fewer and shorter silent pauses [Siegman and Pope, 1965]. Smith et al. [1975] showed that speech rate is positively correlated with perceived competence (conscientiousness). Interestingly, the same authors found that speech rate has an inverted-U relationship with benevolence (agreeableness), suggesting a need for non-linear models.

We added prosodic features based on the audio data of the EAR conversation extracts. As the EAR recorded the participants at anytime of the day, it was necessary to automatically remove any non-voiced signal. We used PRAAT [Boersma, 2001] to compute features characterising the voice's pitch and intensity (mean, extremas and standard deviation), and we added an estimate of the speech rate by dividing the number of words by the voiced time. As an important aspect of this work is that all features are extracted without any manual annotation beyond transcription, we did not filter out utterances from other speakers that may have been captured by the EAR even though it utilised a microphone pointing towards the participant's head. Although advances in speaker recognition techniques might improve the accuracy of prosodic features, we make the assumption that the noise introduced by the surrounding speakers has little effect on our prosodic features, and that it therefore does not affect the performance of the statistical models. This assumption still remains to be tested, as the personality similarity-attraction effect [Byrne and Nelson, 1965] might influence the personality distribution of a participant's conversational partners.

We included all the features mentioned in this section (117) in the models based on the EAR corpus. Models computed using the essays corpus contain only LIWC and MRC features (102), as utterance type features are only meaningful in dialogues.

3.2.3 Correlational analysis

In order to assess what individual features are important for modelling personality regardless of the model used, we report previous correlational studies for the LIWC features on the same data as well as analyses of the new MRC, utterance type and prosodic features. The LIWC features were already analysed by Mehl et al. [2006]

for the EAR dataset, and by Pennebaker and King [1999] for the essays.[1] Tables 3.7 to 3.10 show the features correlating significantly with the personality ratings ($p < .05$, correlations above .05 only). While these studies test multiple hypotheses on the same data, we do not adjust the reported significance levels (e.g. Bonferroni correction), in order to make sure that all significant relations are identified, at the risk of inflating the significance of some of them.[2] This analysis combines together results from previous studies and new findings that provide insight into the features likely to influence the personality recognition models in Sections 3.3.3, 3.4.3 and 3.5.3.

The correlations in Tables 3.7 and 3.8 between LIWC and MRC features and the essays data set show that although extraversion is very well perceived in conversations, it is not strongly reflected through written language, as the correlation magnitudes for the essays dataset are noticeably low. Table 3.9 shows that word count (WC) is a very important feature for modelling extraversion in conversation, both for observer reports and self-reports. Interestingly, this marker does not hold for written language (see Table 3.8). Other markers common to observed and self-reported extraversion include the variation of intensity (Int-stddev), the mean intensity (Int-mean), word repetitions (Unique), words with a high concreteness (Conc) and imagery (Imag). See Table 3.10. On the other hand, words related to anger, affect, swearing, and positive and negative emotions (Posemo and Negemo) are perceived as extravert, but they do not mark self-assessed extraversion in conversations.

Tables 3.9 and 3.10 show that for emotional stability, only a few markers hold for both self-reports and observer reports: a high word count and a low mean pitch (Pitch-mean). Surprisingly, observed emotional stability is associated with swearing and anger words, but not the self-assessed ratings. As reported by Mehl et al. [2006], neurotics are expected to produce more self-references (Self and I). Pennebaker and King [1999] show that the neurotics' use of self-references is also observed in the essays, as well as the use of words related to negative emotions and anxiety. Table 3.10 shows that in conversations, self-assessed neurotics tend

[1] Our correlations differ from Pennebaker and King's study because we use additional student essays collected during the following years.

[2] Pennebaker and King [1999] and Mehl et al. [2006] do not report using any adjustment for multiple significance tests in their studies.

Trait	Extraversion	Emotional stability	Agreeableness	Conscientiousness	Openness to experience
LIWC					
Achieve	.03	.01	-.01	.02	-.07**
Affect	.03	-.07**	-.04	-.06**	.04*
AllPct	-.08**	-.04	-.01	-.04	.10**
Anger	-.03	-.08**	-.16**	-.14**	.06**
Anx	-.01	-.14**	.03	.05*	-.04
Apostro	-.08**	-.04	-.02	-.06**	.05**
Article	-.08**	.11**	-.03	.02	.11**
Assent	.01	.02	.00	-.04	.04*
Body	-.05**	-.04	-.04*	-.04*	.02
Cause	.01	-.03	.00	-.04	-.05*
Certain	.05*	-.01	.03	.04*	.04
Cogmech	-.03	-.02	-.02	-.06**	.02
Comm	-.02	.00	-.01	-.05**	.03
Comma	-.02	.01	-.02	-.01	.10**
Death	-.02	-.04	-.02	-.06**	.05*
Dic	.05*	-.09**	.06**	.06**	-.20**
Excl	-.01	.02	-.02	-.01	.07**
Exclam	.00	-.05*	.06**	.00	-.03
Family	.05*	-.05*	.09**	.04*	-.07**
Feel	-.01	-.09**	.04	.02	-.04*
Fillers	-.04*	.01	-.01	-.03	-.01
Friends	.06**	-.04*	.02	.01	-.12**
Future	-.02	.01	.02	.07**	-.04
Groom	-.02	-.02	.01	.01	-.05**
Hear	-.03	.00	-.01	-.04*	.04*
Home	-.01	-.02	.04*	.06**	-.15**
Humans	.04	-.02	-.03	-.08**	.04
I	.05*	-.15**	.05*	.04	-.14**
Incl	.04*	-.01	.03	.04*	-.03
Inhib	-.03	.02	-.02	-.02	.04*
Insight	-.01	-.01	.00	-.03	.05*
Job	.02	.01	.01	.05**	-.05**
Leisure	-.03	.07**	.03	-.01	-.05**
Metaph	-.01	.01	-.01	-.08**	.08**
Motion	.03	-.01	.05*	.03	-.13**
Music	-.04*	.06**	-.01	-.07**	.10**
Negate	-.08**	-.12**	-.11**	-.07**	.01
Negemo	-.03	-.18**	-.11**	-.11**	.04
Nonfl	-.03	.01	.01	-.05*	.02
Number	-.03	.05*	-.03	-.02	-.06**
Occup	.03	.05*	.04	.09**	-.18**
Optim	.03	.04	.01	.08**	-.07**
Other	.06**	-.01	.03	.01	.01
Othref	.07**	.02	.01	.01	.06**
Parenth	-.06**	.03	-.04*	-.01	.10**
Period	-.05*	-.03	-.01	-.01	.04
Physcal	-.02	-.05*	-.03	-.03	.01

Table 3.7: Pearson's correlation coefficients between LIWC features and personality ratings for the essays dataset, based on the analysis from Pennebaker and King [1999] (* = significant at the $p < .05$ level, ** = $p < .01$). Only features that correlate significantly with at least one trait are shown.

to have a low and constant voice intensity (Int-mean and Int-stddev), while these markers are not used by observers at all.

While emotional stability is expressed differently in various datasets, some markers of agreeableness are consistent: words related to swearing (Swear) and anger (Anger) indicate both self-assessed and observed disagreeableness, regardless of the source of language. See Tables 3.7, 3.8 and 3.9. Interestingly, Table 3.10

Trait	Extraversion	Emotional stability	Agreeableness	Conscientiousness	Openness to experience
LIWC (2)					
Posemo	.07**	.07**	.05*	.02	.02
Posfeel	.07**	-.01	.03	-.02	.08**
Preps	.00	.06**	.04	.08**	-.04
Present	.00	-.12**	-.01	-.03	-.09**
Pronoun	.07**	-.12**	.04*	.02	-.06**
Qmark	-.06**	-.05*	-.04	-.06**	.08**
Quote	-.05*	-.02	-.01	-.03	.09**
Relig	.00	.03	.00	-.06**	.07**
Sad	.00	-.12**	.00	.01	-.01
School	.03	.05**	.06**	.10**	-.20**
See	.00	.09**	.00	-.03	.05**
Self	.07**	-.14**	.06**	.04*	-.14**
Semic	-.03	.02	.02	.00	.05**
Sexual	.07**	-.02	.00	-.04	.09**
Sixltr	-.06**	.06**	-.05*	.02	.10**
Sleep	-.01	-.03	-.02	.03	-.08**
Social	.08**	.00	.02	-.02	.02
Space	-.02	.05*	.03	.01	-.04
Sports	.01	.09**	.02	.00	-.05**
Swear	-.01	.00	-.14**	-.11**	.08**
Tentat	-.06**	-.01	-.03	-.06**	.05*
Time	-.02	.02	.07**	.09**	-.15**
TV	-.04	.04*	-.02	-.04*	.04
Unique	-.05**	.10**	-.04*	-.05*	.09**
Up	.03	.06**	.02	-.01	-.06**
WC	.03	-.06**	.01	.02	.05*
We	.06**	.07**	.04*	.01	.04
WPS	-.01	.02	.02	-.02	.06**
You	-.01	.03	-.06**	-.04*	.11**
MRC					
AOA	-.01	.05*	-.04*	.06**	.11**
Brown-freq	.05*	-.06**	.03	.06**	-.07**
Conc	.02	-.06**	.03	-.01	-.10**
Fam	.08**	-.05*	.08**	.05**	-.17**
Imag	.05*	-.04*	.05*	.00	-.08**
K-F-freq	-.01	.10**	.00	.05*	.07**
K-F-ncats	.06**	-.04*	.08**	.07**	-.12**
K-F-nsamp	.06**	-.01	.03	.05**	-.07**
Meanc	.06**	-.10**	.05**	-.01	-.11**
Meanp	.02	-.02	.05*	.00	-.04*
Nlet	-.09**	.09**	-.03	.00	.15**
Nphon	-.08**	.08**	-.03	.01	.14**
Nsyl	-.07**	.07**	-.02	.04	.13**
T-L-freq	.01	.10**	.01	.06**	.05**

Table 3.8: Continuation of Table 3.7, i.e. Pearson's correlation coefficients between LIWC and MRC features and personality ratings for the essays dataset (* = significant at the $p < .05$ level, ** = $p < .01$). Only features that correlate significantly with at least one trait are shown.

shows that agreeable people do more back-channelling (Prompt), suggesting that they listen more to their conversational partners. While observers do not seem to take prosody into account for evaluating agreeableness, Table 3.10 shows that prosodic cues such as the pitch variation (Pitch-stddev) and the maximum voice intensity (Max-int) indicate self-assessed disagreeableness.

As far as markers of conscientiousness are concerned, Tables 3.7 to 3.9 show that they are similar to those of agreeableness, as unconscientious participants

Dataset	Observer reports					Self-reports				
Trait	**Extra**	**Emot**	**Agree**	**Consc**	**Open**	**Extra**	**Emot**	**Agree**	**Consc**	**Open**
LIWC										
Affect	.40**	.13	-.20	-.24*	.00	.05	-.13	-.17	-.19	.13
Anger	.37**	.30**	-.49**	-.56**	-.14	-.02	.07	-.30**	-.30**	.10
Articles	.21*	.32**	.03	-.15	.14	.03	.00	.04	-.09	-.04
Assent	-.29**	-.02	.30**	.24*	.03	-.11	-.05	.19	-.03	.08
Cause	-.13	-.23*	.03	.15	.00	.00	-.09	.07	-.02	-.23*
Cogmech	.04	-.01	.24*	.20*	.23*	.11	.01	.08	.00	-.06
Comm	-.18	-.27**	-.14	.00	-.26*	-.01	-.13	.20*	.12	-.17
Dic	-.07	-.16	-.17	-.05	-.08	.02	-.15	.16	-.01	-.20*
Discrep	.08	-.03	.13	.10	.23*	.10	-.01	.15	.09	-.09
Eating	.25*	.15	-.31**	-.43**	-.11	-.03	-.02	-.10	-.19	-.05
Family	.26*	-.23*	-.12	-.03	-.04	.14	-.02	.26**	.04	-.14
Feel	.21*	.06	.03	-.03	.05	.08	.05	-.08	.02	.02
Female	.29**	-.03	.04	.03	-.17	.24*	.07	.29**	.12	-.22*
Filler	-.01	-.19	.04	.20*	.01	-.05	-.13	.20	.18	-.08
Friend	.14	-.01	-.08	-.13	-.14	.20*	.01	.05	.16	-.11
Hear	-.20	-.23*	-.19	-.07	-.29**	-.04	-.08	.13	.07	-.19
Home	-.02	-.19	.03	.04	.06	.04	-.12	.29**	-.03	-.07
Humans	-.01	.21*	-.01	-.23*	-.12	.07	-.03	-.20	-.06	.01
I	.03	-.41**	-.21*	-.08	-.17	.21*	-.16	.23*	.01	-.08
Inhib	.19	.01	-.22*	-.14	.00	.02	.02	-.18	-.11	-.12
Insight	.04	-.02	.34**	.29**	.32**	-.06	-.10	.03	.01	.05
Metaph	.30**	.07	-.10	-.26*	-.02	.20	.10	-.10	-.09	.03
Money	-.02	.24*	-.13	-.24*	.01	-.08	.01	-.22*	-.06	-.15
Negemo	.36**	.18	-.44**	-.49**	-.11	.03	-.05	-.16	-.25*	.10
Nonfl	-.01	.05	.09	.24*	.06	-.02	.17	-.03	-.02	.17
Other	.09	.02	-.07	-.09	-.17	.02	.04	.05	.05	-.28**
Othref	.00	.05	-.13	-.14	-.22*	.02	.13	.07	.01	-.19
Past	-.19	-.07	-.25*	-.18	-.31**	-.10	-.18	-.05	.05	-.26**
Physcal	.30**	.24*	-.39**	-.47**	-.17	-.07	-.06	-.16	-.27**	.05
Posfeel	.28**	.04	.05	.14	.05	.06	-.14	-.07	.23*	.11
Pronoun	-.02	-.30**	-.23*	-.17	-.28**	.12	-.07	.19	.05	-.21*
Relig	.30**	.06	-.09	-.27**	-.07	.26*	.15	-.06	-.09	.04
Self	.09	-.42**	-.25*	-.13	-.15	.25*	-.17	.18	.02	-.08
Senses	-.04	-.12	-.18	-.15	-.26*	.03	-.10	.12	.03	-.14
Sexual	.24*	.21*	-.49**	-.48**	-.22*	-.05	.04	-.19	-.23*	.04
Sixltr	-.04	-.04	.25*	.30**	.24*	-.20	-.15	-.01	.19	.03
Social	-.04	-.06	-.17	-.15	-.31**	.06	.04	.12	.06	-.21*
Space	.03	.18	-.21*	-.24*	-.07	-.10	.09	-.18	.01	.23*
Sports	.10	.28**	-.15	-.19	-.11	.03	.21*	-.15	-.05	-.03
Swear	.30**	.27**	-.51**	-.61**	-.17	-.08	.06	-.28**	-.29**	.06
Tentat	-.04	.15	.26*	.15	.30**	-.14	.04	.05	.14	.05
Unique	-.6**	-.18	-.03	-.03	-.12	-.32**	-.22*	-.18	-.05	-.03
Up	.06	.04	-.08	-.11	-.05	.06	.07	-.05	.03	.31**
WC	.63**	.28**	.10	.07	.20	.29**	.22*	.18	.03	.06

Table 3.9: Pearson's correlation coefficients between LIWC features and personality ratings for the EAR dataset, based on the analysis from Mehl et al. [2006] (* = significant at the $p < .05$ level, ** = $p < .01$). Only features that correlate significantly with at least one trait are shown.

also use words related to swearing (Swear), anger (Anger) and negative emotions (Negemo), regardless of the dataset and assessment method. On the other hand, observed conscientiousness is associated with words expressing insight, back-channels (Prompt), longer words (Nphon, Nlet, Nsyl and Sixltr) as well as words that are acquired late by children (AOA), while self-assessed conscientiousness is mostly expressed through positive feelings (Posfeel) in conversations. The avoidance of negative language seems to be the main marker of conscientiousness in essays, as

Dataset	Observer reports					Self-reports				
Trait	Extra	Emot	Agree	Consc	Open	Extra	Emot	Agree	Consc	Open
Prosody										
Int-max	.42**	.12	.07	-.13	.05	.19	.10	-.25*	-.01	.14
Int-mean	.32**	.20	-.02	-.06	.04	.21*	.22*	-.05	-.16	.03
Int-stddev	.40**	.03	-.08	-.12	-.08	.36**	.28**	.00	-.06	.10
Pitch-max	.28**	.10	.13	.05	.23*	-.03	-.11	-.10	-.03	.01
Pitch-mean	.17	-.45**	.06	.04	-.18	.12	-.25*	.07	.03	-.04
Pitch-min	-.17	-.23*	-.02	.08	-.04	.09	-.08	.21*	.04	.08
Pitch-stddev	-.13	.13	.07	.03	.11	-.28**	.01	-.34**	.03	-.03
Voiced	.23*	.27**	.06	.03	.21*	-.02	.07	-.04	-.03	.03
Word-per-sec	.07	-.14	-.12	-.04	-.17	.20*	.07	.09	.02	.04
MRC										
AOA	-.23*	.01	.26**	.26**	.21*	-.12	.04	.05	-.05	.08
Brown-freq	-.26*	-.41**	-.08	.07	-.16	-.04	-.15	.14	.07	-.12
Conc	.24*	-.05	-.20*	-.33**	-.32**	.23*	-.10	.01	-.12	-.02
Fam	-.17	-.28**	-.24*	-.07	-.18	-.03	-.21*	.17	.01	-.13
Imag	.33**	.00	-.23*	-.33**	-.35**	.25*	-.09	.01	-.06	-.03
K-F-freq	-.27**	-.04	.07	.17	.16	-.22*	-.06	-.24*	.05	-.01
K-F-ncats	-.24*	-.24*	-.03	.08	.00	-.01	-.06	.17	.05	-.23*
K-F-nsamp	-.24*	-.20*	-.03	.16	.20	-.15	-.04	.03	.08	-.17
Meanc	.29**	-.10	-.18	-.25*	-.34**	.23*	-.12	.08	-.06	-.07
Nlet	-.14	.17	.25*	.31**	.25*	-.23*	.03	-.18	.13	.12
Nphon	-.12	.09	.25*	.36**	.28**	-.16	.02	-.20	.15	.13
Nsyl	-.16	-.04	.23*	.34**	.19	-.13	-.02	-.06	.12	.10
T-L-freq	-.24*	-.06	.06	.16	.13	-.19	-.07	-.18	.06	-.08
Utterance type										
Assertion	-.05	-.21*	-.03	.01	-.09	-.02	-.06	-.09	.21*	-.14
Command	.00	.01	-.08	-.20*	.00	.13	.21*	-.01	.00	.16
Prompt	-.10	.07	.36**	.27**	.25*	-.05	.01	.22*	-.05	.02
Question	.13	.22*	-.16	-.11	-.04	.01	-.01	-.02	-.24*	.10

Table 3.10: Continuation of Table 3.9, i.e. Pearson's correlation coefficients between features and personality ratings for the EAR dataset (* = significant at the $p < .05$ level, ** = $p < .01$). Only features that correlate significantly with at least one trait are shown.

all other features in Table 3.7 correlate only weakly with the self-reports.

Tables 3.7 and 3.8 show that openness to experience is the trait yielding the highest correlations in the essays corpus: articles, second person pronouns (You) and long words (Sixltr) indicate openness, while non-open participants tend to talk about their occupations (Occup, Home and School) and themselves (Self). As far as conversations are concerned, observers use similar cues for openness as with conscientiousness, such as insight words, longer words, back-channels and a high age of acquisition (AOA).

This section shows that features are likely to vary depending on the source of language and the method of assessment of personality. While such analyses can help evaluate the usefulness of individual features, the question of how such features should be combined to predict personality accurately is addressed by the statistical models.

3.2.4 Statistical models

Various systems require different levels of granularity for modelling personality: it might be more important to cluster users into large groups as correctly as possible, or the system might need to discriminate between individual users. Depending on the application and the adaptation capabilities of the target system, it is possible to use different types of personality models, depending on whether personality modelling is treated as a classification problem, as in previous work by Argamon et al. [2005] and Oberlander and Nowson [2006], or whether we model personality traits via the scalar values actually generated by the self-reports and observer methods used in the corpus collection described in Section 3.2.1.

To support applications in dialogue system adaptation where the output generation is limited to a few points at extremes of a personality scale (such as introvert vs. extravert language or neurotic vs. emotionally stable), we develop classification models by splitting our subjects into two equal size groups.[3]

However, if we model personality traits as scalar values, we have two choices. We can treat personality modelling as a regression problem or as a ranking problem. While regression models can replicate the actual scalar values seen in the personality ratings data, there is also a good argument for treating personality as a ranking problem because by definition, personality evaluation assesses relative differences between individuals, e.g. one person is described as an extravert because the average population is not. Moreover, Freund et al. [1998] argue that ranking models are a better fit to learning problems in which scales have arbitrary values (rather than reflecting real world measures).

For classification and regression models, we use the Weka toolbox [Witten and Frank, 2005] for training and evaluation. In order to evaluate models of personality classification, we compare six different learning algorithms against a baseline returning the majority class. The classification algorithms analysed here are C4.5 decision tree learning (J48), nearest neighbour with one neighbour (NN), Naive Bayes (NB), Ripper rule learning (JRIP), Adaboost with 10 rounds of boosting (ADA) and support vector machines with a linear kernel (SVM).

For regression, we compare five algorithms with a baseline model returning

[3] Depending on the ratings' distribution, a limitation of equal size splits is that the threshold between classes can differ from the neutral rating suggested by personality questionnaires (e.g. 4 out of 7).

the mean personality score in the training data. We focus on a linear regression model (LR), an M5' regression tree (M5R), an M5' model tree returning a linear model (M5), a REPTree decision tree (REP), and a model based on support vector machines with a linear kernel (SVM). Parameters of the algorithms are set to Weka's default values.

Concerning the ranking problem, we train personality models for each Big Five trait using RankBoost, a boosting algorithm for ranking [Freund et al., 1998, Schapire, 1999]. Given a personality trait to model, the linguistic features and personality scores are converted into a training set $\mathcal{T}$ of *ordered pairs* of examples x, y:

$$\begin{aligned}\mathcal{T} = \{(x, y) |\ & x, y \text{ are language samples from two individuals,} \\ & x \text{ has a higher score than } y \text{ for that personality trait}\}\end{aligned}$$

The learnt models are represented by a set of m indicator functions $h_s(x)$ for $1 \leq s \leq m$. The indicator functions are calculated by selecting one of the features described in Section 3.2.2 and by thresholding its value. For example, one indicator function is:

$$h_{100}(x) = \begin{cases} 1 & \text{if } \textsc{Word-per-sec}(x) \geq .73 \\ 0 & \text{otherwise} \end{cases}$$

So $h_{100}(x) = 1$ if x's average speech rate is above .73 words per second. A weight α_s is associated with each indicator function, and the *ranking score* for an example x is calculated as

$$F(x) = \sum_{s=1}^{m} \alpha_s h_s(x)$$

This score is used to rank various language samples (written text or conversation extracts), with the goal of duplicating the ranking found in the training data, and the training examples are used to set the weights α_s. The training algorithm estimates the indicator functions and the weights in order to minimise the following loss function:

$$Loss = \frac{1}{|\mathcal{T}|} \sum_{(x,y) \in \mathcal{T}} smallerThan(F(x), F(y))$$

The *smallerThan* function returns 1 if its first argument is smaller or equal to the second—i.e. the ranking scores of the (x, y) pair are misordered—and 0 otherwise.

In other words, the ranking loss is the percentage of misordered pairs, for which the order of the predicted scores does not match the order dictated by the personality scores from the questionnaire.

Most of the techniques used in this work express the learnt models as rules or decision trees, which support the analysis of differences in the personality models (see Sections 3.3.3, 3.4.3 and 3.5.3).

3.3 Classification results

We evaluate binary classification models based on the essays corpus with self-reports of personality, as well as models based on the EAR corpus with both self and observer reports. All results are averaged over a 10-fold cross-validation, and all significance tests were done using a two-tailed paired t-test at the $p < .05$ level.[4]

3.3.1 Essays corpus

Classification results for the essays corpus with self-reports are in Table 3.11. Interestingly, openness to experience is the easiest trait to model as five classifiers out of six significantly outperform the baseline and four of them produce their best performance for that trait, with accuracies up to 62.1% using support vector machines (SVM). Emotional stability produces the second best performance for four classifiers out of six, with 57.4% accuracy for the SVM model. Conscientiousness is the hardest trait to model as only two classifiers significantly outperform the baseline, however the SVM model performs as well as the best model for extraversion and agreeableness, with 55% correct classifications.

We find that support vector machines generally perform the best, with Naive Bayes and Adaboost in second position. SVM significantly outperforms the majority class baseline for each trait. A J48 decision tree for recognising extraversion is shown in Figure 3.2, and the rule-based JRip model classifying openness to experience with 58.8% accuracy is illustrated in Table 3.15.

Feature set comparison: In order to evaluate how each feature set contributes to the final result, we trained binary classifiers using the algorithms producing the best

[4]Throughout this thesis, we do not adjust the reported significance levels for multiple hypothesis testing.

Trait	Base	J48	NN	NB	JRIP	ADA	SVM
Extraversion	50.04	54.44•	53.27•	53.35•	52.70	**55.00•**	54.93•
Emotional stability	50.08	51.09	51.62	56.42•	55.90•	55.98•	**57.35•**
Agreeableness	50.36	53.51•	50.16	53.88•	52.63	52.71	**55.78•**
Conscientiousness	50.57	51.37	52.10	53.80	52.71	54.45•	**55.29•**
Openness to experience	50.32	54.24•	53.07	59.57•	58.85•	59.09•	**62.11•**

• statistically significant improvement over the majority class baseline (two-tailed paired t-test, $p < .05$)

Table 3.11: Classification accuracy with two equal size bins on the essays corpus, using self-reports. Models are the majority class baseline (Base); J48 decision tree (J48); nearest neighbour (NN); Naive Bayes (NB); JRip rule set (JRIP); Adaboost (ADA); support vector machines (SVM).

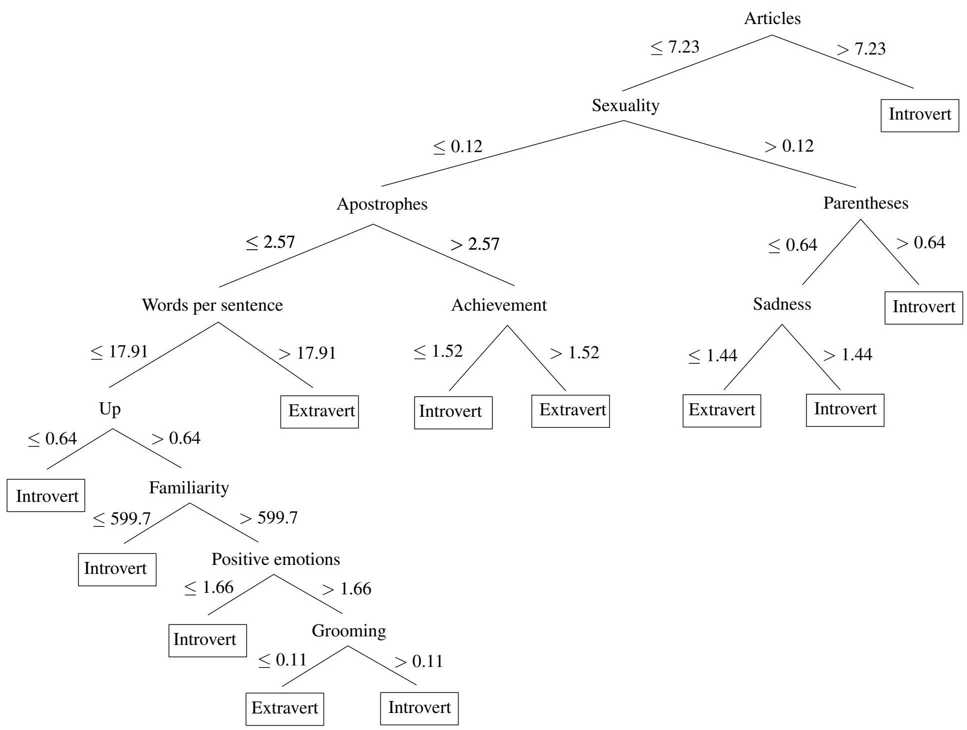


Figure 3.2: J48 decision tree for binary classification of extraversion, based on the essays corpus and self-reports.

overall results with each feature set. We only analyse LIWC and MRC features for the essays corpus, as utterance type and prosodic features do not apply to written texts. We use the Naive Bayes, Adaboost and SVM classifiers as they give the best performance with the full feature set. Results are shown in Table 3.12.

Remarkably, we can see that the LIWC features outperform the MRC features for every trait, and the LIWC features on their own always perform slightly better than the full feature set. This clearly suggests that MRC features are not as helpful as the LIWC features for classifying personality from written text, however Table 3.12 shows that they can still outperform the baseline for four traits out of five.

Concerning the learning algorithms, we find that Adaboost performs the best for extraversion (56.3% correct classifications), while SVM models produce the best accuracy for all other traits. This suggests that support vector machines are promising for modelling personality in general. The easiest trait to model is still openness to experience, with 62.5% accuracy using LIWC features only.

Feature set	**None**	**LIWC features**			**MRC features**		
Classifier	Base	NB	ADA	SVM	NB	ADA	SVM
Set size	0	88	88	88	14	14	14
Extraversion	50.04	52.71	**56.34**•	52.75	52.87•	51.45	53.88
Emotional stability	50.08	56.02•	55.33•	**58.20**•	52.39	52.06	53.52•
Agreeableness	50.36	54.12•	52.71	**56.39**•	53.03•	52.06	53.31•
Conscientiousness	50.57	53.92•	54.48•	**55.62**•	53.03	52.95	53.84
Openness to experience	50.32	58.92•	58.64•	**62.52**•	55.41•	56.70•	57.47•

• statistically significant improvement over the majority class baseline (two-tailed paired t-test, $p < .05$)

Table 3.12: Classification accuracies with two equal size bins on the essays corpus using the majority class baseline (Base),[5] Naive Bayes (NB), Adaboost (ADA) and support vector machine (SVM) classifiers, for different feature sets. Best model for each trait are in bold.

3.3.2 EAR corpus

Classification accuracies for the EAR corpus are in Table 3.13. We find that extraversion is the easiest trait to model using observer reports, with both Naive Bayes and Adaboost outperforming the baseline with an accuracy of 73.0%. The J48 decision tree for extraversion with a 66.8% accuracy is shown in Figure 3.3. Emotional stability is modelled with comparable success using a Naive Bayes classifier, however the improvement over the baseline is lower than with extraversion (22.8% vs. 25.2%) and other classifiers do not perform as well. Models of observed conscientiousness also outperform the baseline, with 67.7% accuracy using a Naive Bayes classifier, while the best model for agreeableness produces 61.3% correct classifications. None of the models for openness to experience significantly outperform the baseline, which suggests that openness to experience is expressed more clearly in

[5]Although equal size bins were used, the baseline accuracies differ from 50% because of the random sampling of the cross-validation.

stream-of-consciousness essays and self-reports than in the EAR dataset. Support vector machines do not perform as well as with the essays corpus, probably because of the sparseness of the dataset. Self-reports are much harder to model than observer reports given the same dataset size, as none of the self-report classifiers significantly outperform the majority class baseline.

Data	Trait	Base	J48	NN	NB	JRIP	ADA	SVM
Obs	Extra	47.78	66.78	59.33	**73.00**•	60.44	**73.00**•	65.78
Obs	Emot	51.11	62.56	58.22	**73.89**•	56.22	48.78	60.33
Obs	Agree	47.78	48.78	51.89	**61.33**•	51.89	52.89	56.33
Obs	Consc	47.78	57.67	61.56	**67.67**•	61.56	60.22 •	57.11
Obs	Open	47.78	52.22	46.78	**57.00**	49.67	50.56	55.89
Self	Extra	47.78	48.78	49.67	**57.33**	50.56	54.44	49.89
Self	Emot	**51.11**	45.56	46.78	50.44	46.78	41.89	44.33
Self	Agree	52.22	47.89	50.89	**58.33**	56.89	55.22	52.33
Self	Consc	51.11	33.44	45.56	39.33	43.11	46.11	**53.22**
Self	Open	51.11	52.00	42.22	**61.44**	45.00	56.00	47.78

• statistically significant improvement over the majority class baseline (two-tailed paired t-test, $p < .05$)

Table 3.13: Classification accuracy with two equal size bins on the EAR corpus, for observer ratings (Obs) and self-reports (Self). Models are majority class baseline (Base); J48 decision tree (J48); nearest neighbour (NN); Naive Bayes (NB); JRip rules set (JRIP); Adaboost (ADA); support vector machines (SVM).

Feature set	None	Type	LIWC	MRC	Prosody
Set size	0	4	88	14	11
Extraversion	47.78	45.67	**68.89**•	68.78•	67.56•
Emotional stability	51.11	60.22	**69.89**•	60.78	61.78
Agreeableness	47.78	57.56	54.00	**58.67**	50.44
Conscientiousness	47.78	59.67	60.22	**66.78**•	52.11
Openness to experience	47.78	53.11	61.11	54.00	**64.56**•

• statistically significant improvement over the majority class baseline (two-tailed paired t-test, $p < .05$)

Table 3.14: Classification accuracies for the EAR corpus with observer reports using the Naive Bayes classifier, for different feature sets (None=baseline, Type=utterance type). Best models are in bold.

Feature set comparison: For the EAR corpus we investigate the importance of all four feature sets: utterance type, LIWC, MRC and prosodic features. We use the Naive Bayes models with the observer ratings as they perform the best with all features. Interestingly, Table 3.14 shows that the high classification accuracies

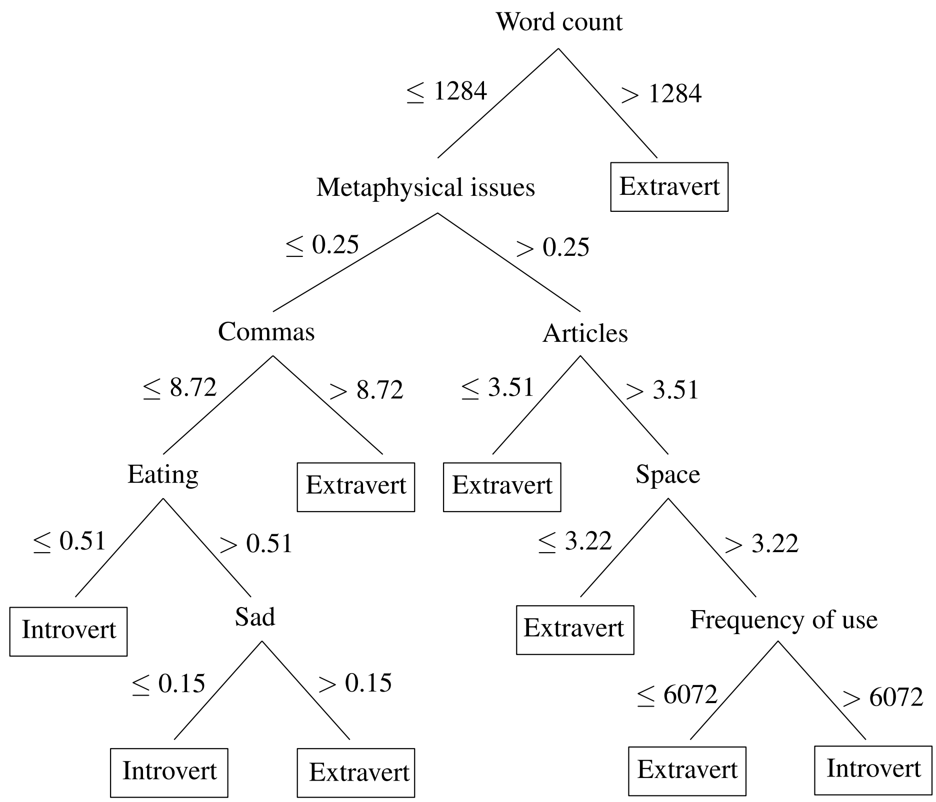


Figure 3.3: J48 decision tree for binary classification of extraversion, based on the EAR corpus and observer reports.

for extraversion come from a combination of LIWC, MRC and prosodic features, as they all outperform the baseline on their own, but do not do as well as the 73.0% accuracy with the full feature set. LIWC features are the main indicators of emotional stability, although the model with all features still performs better. MRC features are the most important for classifying conscientiousness (66.8%), while prosodic features produce the best model of openness to experience with 64.6% accuracy, improving on the model with all features. Although utterance type features never outperform the baseline on their own, the lack of significance could be the result of the small dataset size, since Section 3.2.3 showed that some utterance type features strongly correlate with several personality traits.

3.3.3 Qualitative analysis

Decision trees and rule-based models can be easily understood, and can therefore help to uncover new linguistic markers of personality. Our models replicate previous findings, such as the link between verbosity and extraversion (c.f. *Word count* node of Figure 3.3), but they also provide many new markers.

#	**Ordered rules**
1	(School $\geq$ 1.47) and (Motion $\geq$ 1.71) $\Rightarrow$ NOT OPEN
2	(Occup $\geq$ 2.49) and (Sixltr $\leq$ 13.11) and (School $\geq$ 1.9) and (I $\geq$ 10.5) $\Rightarrow$ NOT OPEN
3	(Fam $\geq$ 600.33) and (Friends $\geq$ 0.67) $\Rightarrow$ NOT OPEN
4	(Nlet $\leq$ 3.50) and (Number $\geq$ 1.13) $\Rightarrow$ NOT OPEN
5	(School $\geq$ 0.98) and (You $\leq$ 0) and (AllPct $\leq$ 13.4) $\Rightarrow$ NOT OPEN
6	Any other feature values $\Rightarrow$ OPEN

Table 3.15: JRip rule set for binary classification of openness to experience, based on the essays corpus.

The model of self-assessed openness to experience detailed in Table 3.15 shows that students referring a lot to school work tend to have low scores for that trait (Rules 1, 2 and 5). As expected, the avoidance of longer words is also indicative of a lack of openness (Rules 4 and 5), as well as the use of high-familiarity words and references to friends (Rule 3).

The model of observed extraversion in Figure 3.3 shows that word count is the most important feature for classifying that trait as an observer. The model also suggests that given a low verbosity, extraversion can still manifest itself through the use of words related to metaphysical issues together with few articles, as well as through the use of many commas. The association between extraversion and the avoidance of articles probably reflects the use of more pronouns over common nouns and confirms previous findings associating extraversion with implicit language [Heylighen and Dewaele, 2002].

Interestingly, the decision tree trained on the essays corpus in Figure 3.2 for self-reported extraversion differs a lot from the observer model in Figure 3.3. While word count is the most important feature for observers, it does not seem to be a marker of self-assessed extraversion (see Section 3.2.3), although the number of words per sentence is used to discriminate on a subset of the data. On the other hand, the self-report model associates introversion with the use of articles, which is also the case in the observer model. While sexual content does not affect the observer model, it is the second most important feature for modelling self-reported extraversion. For example, participants using many sex-related words are modelled as introvert, unless they avoid parentheses and words related to sadness.

3.4 Regression results

We also trained regression models using the same corpora. The baseline is a model returning the mean of all personality scores in the training set. We use the relative absolute error for evaluation, which is the ratio between the model's prediction error and the error produced by the baseline. A low relative error therefore indicates that the model performs better than the constant mean baseline, while a 100% relative error implies a performance equivalent to that baseline. All results are averaged over a 10-fold cross-validation, and all significance tests were done using a two-tailed paired t-test at the $p < .05$ level.

3.4.1 Essays corpus

Regression results with the essays corpus and self-reports are in Table 3.16. Paired t-tests show that emotional stability and openness to experience produce models that significantly improve over the baseline. As with the classification task, openness to experience is the easiest trait to model using essays: four regression models out of five outperform the baseline. The M5' model tree produces the best result with a 93.3% relative error for openness to experience (6.7% error decrease), and a 96.4% relative error for emotional stability.

In terms of correlation between the model predictions and the actual ratings, the model for emotional stability and openness to experience produce Pearson's correlation coefficients of .24 and .33, respectively. Although the magnitude of the improvement seems relatively small, one needs to keep in mind the difficulty of the regression task over the binary classification task: it is the most fine-grained personality recognition problem, requiring the association of an exact scalar value with each individual.

Feature set comparison: Table 3.17 provides results for a comparison of LIWC with the MRC feature sets using the linear regression model, the M5' model tree and the support vector machine algorithm for regression. Overall, LIWC features perform better than MRC features except for extraversion, for which the linear regression model with MRC features produces better results than with the full feature set. LIWC features on their own generally perform better than the full feature set

Trait	**Base**	**LR**	**M5R**	**M5**	**REP**	**SVM**
Relative error:						
Extraversion	100.0	**99.17**	99.31	99.22	99.98	100.65
Emotional stability	100.0	96.87•	99.75	**96.43•**	99.35	98.35
Agreeableness	100.0	**98.92**	99.86	99.22	99.78	100.28
Conscientiousness	100.0	98.68	100.62	**98.56**	100.47	99.30
Openness to experience	100.0	93.58•	97.68•	**93.27•**	99.82	94.19•
Correlation:						
Extraversion	0.00	**0.16•**	0.11•	0.16•	0.07•	0.14•
Emotional stability	0.00	0.23•	0.16•	**0.24•**	0.15•	0.22•
Agreeableness	0.00	**0.17•**	0.12•	0.17•	0.07•	0.15•
Conscientiousness	0.00	**0.18•**	0.07•	0.18•	0.04•	0.16•
Openness to experience	0.00	0.32•	0.22•	**0.33•**	0.16•	0.32•

• statistically significant improvement over the mean value baseline (two-tailed paired t-test, $p < .05$)

Table 3.16: Relative error and correlation coefficients for regression models trained on the essays corpus with all features. Models are the mean value baseline (Base), linear regression (LR); M5' regression tree (M5R), M5' model tree with linear models (M5), REPTree (REP) and support vector machines for regression (SVM).

(except for openness to experience), and almost always significantly outperform the baseline. The model for openness to experience produces the lowest relative error, with 6.50% improvement over the baseline.

Feature set	**None**	**LIWC features**			**MRC features**		
Regression model	Base	LR	M5	SVM	LR	M5	SVM
Extraversion	100.0	99.39	99.25•	100.8	**98.79•**	**98.79•**	99.13•
Emotional stability	100.0	96.71•	**96.42•**	98.03	99.49	99.54	99.89
Agreeableness	100.0	**98.50•**	98.52•	99.52	99.75	99.81	99.31•
Conscientiousness	100.0	98.23•	**98.14•**	99.46	99.23	99.23	99.16•
Openness to experience	100.0	**93.50•**	93.70•	94.14•	97.44•	97.44•	97.26•

• statistically significant improvement over the mean value baseline (two-tailed paired t-test, $p < .05$)

Table 3.17: Relative error for regression models trained on the essays corpus with the MRC and LIWC feature sets only. Models are linear regression (LR); M5' model tree (M5); support vector machines for regression (SVM). Best models are in bold.

3.4.2 EAR corpus

Regression results for the EAR corpus are in Table 3.18. A paired t-test (two-tailed, $p < .05$) over the cross-validation folds shows that the error reduction is significant for observed extraversion (79.9% relative error, i.e. 20.1% error decrease), conscientiousness (17.3% improvement) and emotional stability (13.3% improvement). While extraversion is the easiest trait to model from observer ratings, models of agreeableness and openness to experience do not outperform the baseline.

In terms of correlation between the model predictions and the actual ratings, the models for extraversion, emotional stability and conscientiousness respectively produce Pearson's correlation coefficients of .54, .47 and .54, significantly outperforming the baseline. Such correlations are relatively high, given that the average correlation between the ratings of each pair of observers is .54 for extraversion, .29 for emotional stability and .51 for conscientiousness (18 observers, between 31 and 33 data points for each pair).

Linear regression and support vector machines models perform poorly, suggesting that they require a larger dataset as in the essays corpus. As in the classification task, self-reports of the EAR corpus are clearly difficult to model: none of the models show significant improvement over the baseline.

Data	Trait	Base	LR	M5R	M5	REP	SVM
Relative error:							
Obs	Extraversion	100.0	179.16	82.16•	80.15	**79.94**•	140.05
Obs	Emotional stability	100.0	302.74	92.03•	**86.75**•	100.51	162.05
Obs	Agreeableness	100.0	242.68	**96.73**	111.16	99.37	173.97
Obs	Conscientiousness	100.0	188.18	**82.68**•	90.85	98.08	131.75
Obs	Openness to experience	100.0	333.65	101.64	119.53	102.76	213.20
Self	Extraversion	100.0	204.96	104.50	118.44	**99.94**	176.51
Self	Emotional stability	100.0	321.97	104.10	108.39	**99.91**	233.19
Self	Agreeableness	**100.0**	349.87	106.90	110.84	101.64	201.80
Self	Conscientiousness	**100.0**	177.12	103.39	120.29	107.33	124.91
Self	Openness to experience	**100.0**	413.70	107.12	122.68	126.31	233.01
Correlation:							
Obs	Extraversion	0.00	0.24•	0.51•	0.47•	**0.54**•	0.23•
Obs	Emotional stability	0.00	0.15	0.36•	**0.47**•	0.19•	0.21
Obs	Agreeableness	0.00	0.08	0.36•	**0.44**•	0.08	0.33•
Obs	Conscientiousness	0.00	0.28•	**0.54**•	0.48•	0.20	0.39•
Obs	Openness to experience	0.00	0.17	0.02	**0.20**	0.13	0.12
Self	Extraversion	0.00	**0.10**	0.01	0.05	0.00	-0.05
Self	Emotional stability	0.00	-0.10	-0.05	**0.21**	-0.01	-0.11
Self	Agreeableness	0.00	-0.02	-0.10	**0.05**	-0.07	-0.05
Self	Conscientiousness	0.00	0.20	0.04	-0.06	0.03	**0.35**•
Self	Openness to experience	**0.00**	-0.06	-0.05	-0.16	0.00	-0.20

• statistically significant improvement over the mean value baseline (two-tailed paired t-test, $p < .05$)

Table 3.18: Relative error and correlation coefficients for regression models, with observer ratings (Obs) and self-reports (Self) of the EAR corpus. Models are the mean value baseline (Base); linear regression (LR); M5' regression tree (M5R); M5' model tree with linear models (M5); REPTree decision tree (REP); support vector machines for regression (SVM). The relative error of the baseline model is 100%.

Feature set comparison: We trained regression models with each individual feature set using only observer reports, since self-reports did not produce any significant result using all features. We only focus on the three regression tree algorithms as they perform the best with all features. Table 3.19 shows that LIWC are good predictors of observed extraversion, as the REPTree outperforms the same model with all features with a 76.4% relative error (23.6% improvement over the baseline). LIWC features also produce the best regression model for conscientiousness (82.1% relative error, 17.9% improvement). Surprisingly, the best model of emotional stability contains only prosodic features, with a 85.3% relative error (14.7% improvement). This finding suggests that speech cues are crucial for the perception of neuroticism, which could explain why Gill and Oberlander [2003] reported a low correlation between self-assessed and observed emotional stability using text only. As in the classification task, utterance type features do not show any significant improvement on their own.

Set	**Utterance type**			**LIWC features**		
Model	M5R	M5	REP	M5R	M5	REP
Extraversion	100.0	103.7	101.8	81.61	77.84•	**76.38•**
Emotional stability	102.5	103.0	102.6	90.79•	109.6	109.6
Agreeableness	102.4	102.7	111.1	98.49	111.7	102.5
Conscientiousness	100.0	95.04	104.1	**82.13•**	96.62	93.50
Openness to experience	101.1	99.03	109.9	105.1	129.5	103.7
Set	**MRC features**			**Prosodic features**		
Model	M5R	M5	REP	M5R	M5	REP
Extraversion	99.23	102.2	99.69	94.07	90.91	88.31•
Emotional stability	93.13•	96.08	104.4	92.24•	**85.32•**	97.95
Agreeableness	104.1	112.5	102.2	100.0	108.4	108.9
Conscientiousness	97.00	102.0	91.24•	100.0	104.7	101.7
Openness to experience	106.2	111.6	105.5	100.1	113.5	99.93

• statistically significant improvement over the mean value baseline (two-tailed paired t-test, $p < .05$)

Table 3.19: Relative error for regression models trained on the EAR corpus with individual feature sets. Models are M5' regression tree (M5R); M5' model tree with linear models (M5); REPTree regression tree (REP). Best models are in bold.

3.4.3 Qualitative analysis

Regression trees for extraversion and conscientiousness are in Figures 3.4 and 3.5. As suggested by the correlations in Section 3.2.3, the model in Figure 3.4 shows that the voice's pitch and variation of intensity play an important role when modelling extraversion. A high verbal output is perceived as a sign of extraversion (see *Word count* nodes), confirming previous findings [Scherer, 1979]. On the other hand, a low mean pitch combined with a constant voice intensity characterises high introverts.

Figure 3.5 suggests that conscientious people use fewer swear words and content related to sexuality, while preferring longer words. The same figure also shows that conscientious people use fewer pronouns (i.e. a more explicit style), as well as more words related to communication (e.g. *talk*, *share*).

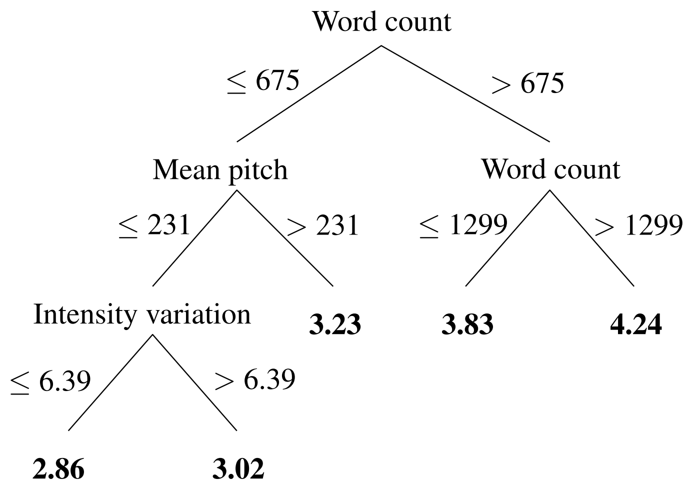


Figure 3.4: M5' regression tree for observed extraversion, computed using the EAR corpus. The target output ranges from 1 to 5.5, as 5.5 is the highest value in the means of the observer ratings. The mean pitch value is expressed in Hertz, and the intensity variation (standard deviation) in decibels.

3.5 Ranking results

Results using RankBoost with both corpora and different feature sets are in Tables 3.20 and 3.21. The models are trained over 100 rounds of boosting. The baseline model ranks the texts randomly, producing a ranking loss of 0.5 on average (lower is better). Results are averaged over a 10-fold cross-validation, and all significance tests were done using a two-tailed paired t-test at the $p < .05$ level.

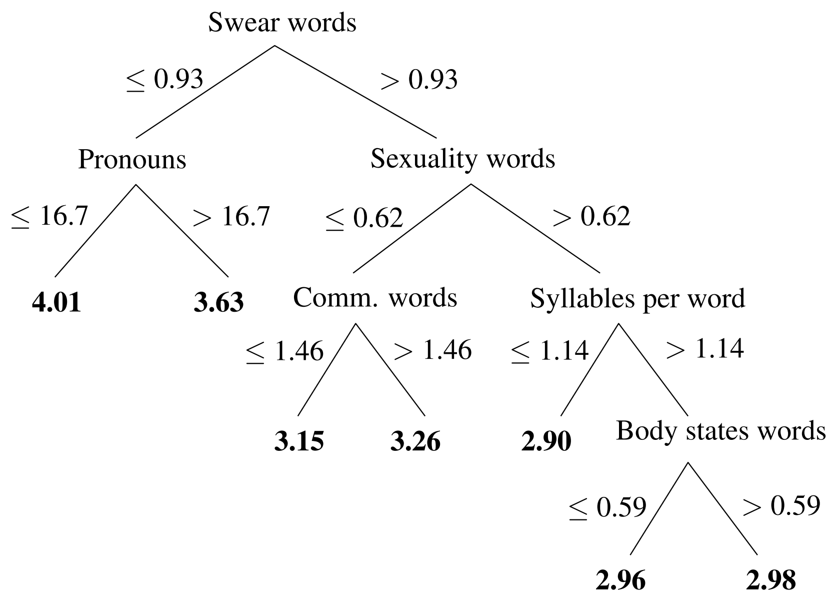


Figure 3.5: M5' regression tree for observed conscientiousness, computed using the EAR corpus. The target output ranges from 1 to 7, where 7 means strongly conscientious (*Comm. words* is the ratio of words related to communication).

3.5.1 Essays corpus

Table 3.20 shows that openness to experience produces the best ranking model with the essays corpus, producing a ranking loss of 0.39. Remarkably, this trait was the easiest to model for all three recognition tasks with that corpus. As it is not the case with conversational data, it seems that stream-of-consciousness essays, or more generally personal writings, are likely to exhibit cues relative to the author's openness to experience. Emotional stability produces the second best model with a ranking loss of 0.42, followed by conscientiousness and extraversion, while the model for agreeableness produces the highest ranking loss. All models significantly outperform the random ranking baseline, but the actual improvement is relatively small.

Feature set comparison: To evaluate which features contribute to ranking accuracy, we trained a ranking model with each feature set. Table 3.20 clearly shows that the LIWC features are the only contributors to model accuracy, as the inclusion of MRC features does not reduce the ranking loss for any trait.

Feature set	Base	All	LIWC	MRC
Extraversion	0.50	0.44•	**0.44•**	0.46•
Emotional stability	0.50	0.42•	**0.42•**	0.47•
Agreeableness	0.50	0.46•	**0.46•**	0.48•
Conscientiousness	0.50	0.44•	**0.44•**	0.47•
Openness to experience	0.50	0.39•	**0.39•**	0.44•

• statistically significant improvement over
the random ordering baseline
(two-tailed paired t-test, $p < .05$)

Table 3.20: Ranking loss for the essays corpus over a 10-fold cross-validation for different feature sets and the random ordering baseline (Base). Best models are in bold (lower is better).

3.5.2 EAR corpus

Concerning the EAR corpus, Table 3.21 shows that models of extraversion, agreeableness, conscientiousness and openness to experience perform better than the random ranking baseline using all features. Emotional stability is the most difficult trait to model, while agreeableness and conscientiousness produce the best results, with ranking losses of 0.31 and 0.33 respectively.

Feature set	None	All	LIWC	MRC	Type	Prosody
Extraversion	0.50	0.35•	0.36•	0.45	0.55	**0.26•**
Emotional stability	0.50	0.41	0.41	**0.39•**	0.43	0.45
Agreeableness	0.50	**0.31•**	0.32•	0.44	0.45	0.54
Conscientiousness	0.50	**0.33•**	0.36•	0.41•	0.44	0.55
Openness to experience	0.50	0.38•	**0.37•**	0.41	0.49	0.44

• statistically significant improvement over the random
ordering baseline (two-tailed paired t-test, $p < .05$)

Table 3.21: Ranking loss for the EAR corpus and observer reports[6] over a 10-fold cross-validation for different feature sets (None=baseline, Type=utterance type). Best models are in bold (lower is better).

Feature set comparison: When looking at individual feature sets, Table 3.21 shows that LIWC features perform significantly better than the baseline for all dimensions but emotional stability, while emotional stability is best predicted by MRC features only (0.39 ranking loss). Interestingly, prosodic features are very good predictors of extraversion, with a lower ranking error than with the full feature set (0.26). This model produces the best overall result, with a 74% chance that the

[6]We also built models of self-reports of personality based on the EAR corpus, but none of them significantly outperforms the baseline.

model will detect the most extravert among any two unseen conversation extracts. As in the previous recognition tasks, utterance type features on their own never significantly outperform the baseline.

3.5.3 Qualitative analysis

The RankBoost rules indicate the impact of each feature on the recognition of a personality trait by the magnitude of the parameter α associated with that feature. Tables 3.22 to 3.24 show the rules with the most impact on the best models for three traits, with the associated α weights. The feature labels are in Table 3.5. For example, the model of extraversion in Table 3.22 confirms previous findings by associating this trait with longer conversations (Rule 5), a high speech rate (Rules 1 and 4) and a high pitch (Rules 2, 6 and 7) [Nass and Lee, 2001]. But new markers emerge, such as a high pitch variation for introverts (Rules 15, 18 and 20), contradicting previous findings reported by Scherer [1979].

Extraversion model with prosodic features					
#	**Positive rules**	α	#	**Negative rules**	α
1	Word-per-sec $\geq$ 0.73	1.43	11	Pitch-max $\geq$ 636.35	-0.05
2	Pitch-mean $\geq$ 194.61	0.41	12	Pitch-slope $\geq$ 312.67	-0.06
3	Voiced $\geq$ 647.35	0.41	13	Int-min $\geq$ 54.30	-0.06
4	Word-per-sec $\geq$ 2.22	0.36	14	Word-per-sec $\geq$ 1.69	-0.06
5	Voiced $\geq$ 442.95	0.31	15	Pitch-stddev $\geq$ 115.49	-0.06
6	Pitch-max $\geq$ 599.88	0.30	16	Pitch-max $\geq$ 637.27	-0.06
7	Pitch-mean $\geq$ 238.99	0.26	17	Pitch-slope $\geq$ 260.51	-0.12
8	Int-stddev $\geq$ 6.96	0.24	18	Pitch-stddev $\geq$ 118.10	-0.15
9	Int-max $\geq$ 85.87	0.24	19	Int-stddev $\geq$ 6.30	-0.18
10	Voiced $\geq$ 132.35	0.23	20	Pitch-stddev $\geq$ 119.73	-0.47

Table 3.22: Subset of the RankBoost model for extraversion with prosodic features only, based on EAR conversations and observer reports. Rows 1-10 represent the rules producing the highest score increase, while rows 11-20 indicate evidence for the other end of the scale, i.e. introversion.

Concerning agreeableness, Rules 1 and 20 in Table 3.23 suggest that agreeable people use longer words but shorter sentences, and Rules 2 and 4 show that they express more tentativity (with words like *maybe* or *perhaps*) and positive emotions (e.g. *happy*, *good*). Anger and swear words greatly reduce the agreeableness score (Rules 12, 13, 18 and 19), as well as the use of negations (Rule 15).

Table 3.24 shows that conscientious people talk a lot about their work (Rule 1), while unconscientious people swear a lot (Rules 11 and 19). Insight words (e.g. *think*, *know*) are also good indicators of conscientiousness, as well as words

Agreeableness model with all features					
#	**Positive rules**	α	#	**Negative rules**	α
1	Nphon $\geq$ 2.66	0.56	11	Fam $\geq$ 601.61	-0.16
2	Tentat $\geq$ 2.83	0.50	12	Swear $\geq$ 0.41	-0.18
3	Colon $\geq$ 0.03	0.41	13	Anger $\geq$ 0.92	-0.19
4	Posemo $\geq$ 2.67	0.32	14	Time $\geq$ 3.71	-0.20
5	Voiced $\geq$ 584	0.32	15	Negate $\geq$ 3.52	-0.20
6	Relig $\geq$ 0.43	0.27	16	Fillers $\geq$ 0.54	-0.22
7	Insight $\geq$ 2.09	0.25	17	Time $\geq$ 3.69	-0.23
8	Prompt $\geq$ 0.06	0.25	18	Swear $\geq$ 0.61	-0.27
9	Comma $\geq$ 4.60	0.23	19	Swear $\geq$ 0.45	-0.27
10	Money $\geq$ 0.38	0.20	20	WPS $\geq$ 6.13	-0.45

Table 3.23: Best RankBoost model based on EAR conversations for agreeableness. Rows 1-10 represent the rules producing the highest score increase, while rows 11-20 indicate evidence for the other end of the scale, i.e. disagreeableness.

expressing positive feelings like *happy* and *love* (Rules 2 and 3). Interestingly, conscientious speakers are modelled as having a high variation of their voice intensity (Rule 4). On the other hand, Rule 20 shows that speaking very loud produces the opposite effect, as well as having a high pitch (Rule 13). Long utterances are also indicative of a low conscientiousness (Rule 12).

The rule sets presented here contain only the most extreme rules of our ranking models, which contain many additional personality cues that are not identified through a typical correlational analysis. For example, a high speech rate and a high mean pitch contribute to a high extraversion ranking in Table 3.22's model, but they do not correlate significantly with observer ratings, as detailed in Table 3.10. Similarly, positive emotion words (Posemo) and the avoidance of long utterances (WPS) indicate agreeableness in the model in Table 3.23, while these features do not correlate significantly with agreeableness ratings.

3.6 Discrete personality modelling in related work

To our knowledge, there are only two other studies on the automatic recognition of personality [Argamon et al., 2005, Oberlander and Nowson, 2006]. Both of these studies have focused on the classification of written texts based on self-reports, rather than using continuous regression and ranking models as we do here.

Argamon et al. [2005] use the essays corpus of Pennebaker and King [1999], so their results are directly comparable to those presented here. As in this thesis, they use a top-down approach to feature definition: their feature set consists of relative

Conscientiousness model with all features					
#	**Positive rules**	α	#	**Negative rules**	α
1	Occup $\geq$ 1.21	0.37	11	Swear $\geq$ 0.20	-0.18
2	Insight $\geq$ 2.15	0.36	12	WPS $\geq$ 6.25	-0.19
3	Posfeel $\geq$ 0.30	0.30	13	Pitch-mean $\geq$ 229	-0.20
4	Int-stddev $\geq$ 7.83	0.29	14	Othref $\geq$ 7.64	-0.20
5	Nlet $\geq$ 3.29	0.27	15	Humans $\geq$ 0.83	-0.21
6	Comm $\geq$ 1.20	0.26	16	Swear $\geq$ 0.93	-0.21
7	Nphon $\geq$ 2.66	0.25	17	Swear $\geq$ 0.17	-0.24
8	Nphon $\geq$ 2.67	0.22	18	Relig $\geq$ 0.32	-0.27
9	Nphon $\geq$ 2.76	0.20	19	Swear $\geq$ 0.65	-0.31
10	K-F-nsamp $\geq$ 329	0.19	20	Int-max $\geq$ 86.84	-0.50

Table 3.24: Best RankBoost model based on EAR conversations for conscientiousness. Rows 1-10 represent the rules producing the highest score increase, while rows 11-20 indicate evidence for the other end of the scale, i.e. unconscientiousness.

frequencies of function words and word categories based on networks of the theory of systemic functional grammar. However, they simplify the task by removing the middle third of the dataset. They train SVM models on the top third and lower third of the essays corpus for extraversion and emotional stability, achieving accuracies of 58% for both traits on this subset of the data.

Oberlander and Nowson [2006] follow a bottom-up feature discovery method by training Naive Bayes and SVM models for four of the Big Five traits on a corpus of personal weblogs, using n-gram features extracted from the dataset. When testing whether the Naive Bayes models generalise to a different corpus of weblogs, Nowson and Oberlander [2007] report binary classification accuracies ranging from 55% for extraversion to 65% for conscientiousness. Interestingly, models trained on the most extreme instances of the original corpus seem to outperform models trained on the full corpus, although no level of significance is mentioned. These studies show that n-grams can be appropriate to model self-reports of personality, although, as Oberlander and Nowson point out, such features are likely to overfit. It would therefore be interesting to test whether the feature sets used here generalise to another dataset. Chapter 8 investigates this issue by applying our personality recognition models to a set of generated utterances, in order to control the personality conveyed by a dialogue system presenting information to the user (see Section 8.4).

Oberlander and Nowson [2006] also report results for three-way and five-way classification, in order to approximate the finer-grained continuous personality ratings used in psychology. They obtain a maximum of 44.7% correct classifications for extraversion with five bins, using raw n-grams (baseline is 33.8%). When keeping only n-grams that correlate with the personality ratings over the full dataset, the highest accuracy increases up to 69.8% for agreeableness. These results are not directly comparable to those presented in this chapter because they are on a different corpus, with different feature sets. Moreover, we have not provided results on such multiple classification experiments, because such models do not take into account the fact that the different classes are part of a total ordering, and thus the resulting models are forced to ignore the importance of features that correlate with that ordering across all classes. We believe that regression and ranking models are more accurate for finer-grained personality recognition (see Sections 3.4 and 3.5).

To evaluate this claim, we first mapped the output of the best classifier to a ranking and compared it with the RankBoost models. We trained a Naive Bayes classifier on the EAR corpus with observer reports and all features, using five equal size bins.[7] For each test fold of a 10-fold cross-validation, we computed the ranking loss produced by the classifier based on the ordering of the five classes. Results in Table 3.25 show that RankBoost significantly outperforms the classifier for four traits out of five ($p < .05$), with an improvement close to significance for emotional stability ($p = 0.12$).

Task	**Ranking**			**Classification**		
Model	**Base**	**NB**	**Rank**	**Base**	**NB**	**Rank**
Extraversion	0.50	0.48	0.35$\bullet$	20.0	32.3	32.1
Emotional stability	0.50	0.50	0.41	20.0	21.9	21.9
Agreeableness	0.50	0.50	0.31$\bullet$	20.0	28.4	37.8
Conscientiousness	0.50	0.46	0.33$\bullet$	20.0	34.7	30.3
Openness to experience	0.50	0.53	0.38$\bullet$	20.0	19.8	26.8

$\bullet$ statistically significant improvement over the other model (two-tailed t-test, $p < .05$)

Table 3.25: Comparison between ranking (Rank) and classification models (NB) for both personality ranking and classification tasks (five bins). Evaluation metrics are ranking loss (lower is better) and classification accuracy (higher is better), respectively. Results are averaged over a 10-fold cross-validation.

Because RankBoost's goal is to minimise the ranking loss, this comparison is likely to favour ranking models. Therefore, we also mapped the output of the Rank-

[7] Oberlander and Nowson use unequal bins defined for each personality trait using standard deviation from the mean.

Boost models to five classification bins to see whether RankBoost could perform as well as a classifier for the classification task. We divided the output ranking into 5 bins, each containing a 20% slice of contiguously ranked instances. We note that this use of ranking is not strictly correct as a classification method, in terms of how the classification task is traditionally conceived. In particular, the ranking method requires all test instances to be supplied together, exploiting the distribution of classes within the test set to achieve classification, whereas a standard classifier is able to classify each test instance without sight of any other test instances.

Results in the *classification* column of Table 3.25 show that the Naive Bayes classifier never outperforms RankBoost significantly, while the ranking model produces a better mean accuracy for agreeableness (38%) and openness to experience (27%), and the same accuracy for emotional stability (22%). In sum, ranking models perform as well for classification and better for ranking compared with our best classifier, thus modelling personality using continuous models is more accurate.

3.7 Discussion and summary

We show that personality can be recognised by computers through language cues.[8] To our knowledge, the results presented here are the first to demonstrate statistically significant results for recognising personality in conversation. We present the first results applying regression and ranking models in order to model personality recognition using the continuous scales traditional in psychology. Different feature sets are also systematically examined.

Table 3.26 summarises results for all the personality traits and recognition tasks analysed in this chapter. What clearly emerges is that extraversion is the easiest trait to model from spoken language, followed by emotional stability and conscientiousness. Concerning written language, models of openness to experience produce the best results for all recognition tasks. Feature selection is important, as some of the best models only contain a small subset of the full feature set. Prosodic features are important for modelling observed extraversion, emotional stability and openness to experience. MRC features are useful for models of emotional stability, while LIWC features are beneficial for all traits. We also analyse qualitatively which fea-

[8] An online demo and a personality recognition tool based on the models presented in this chapter can be downloaded from `www.dcs.shef.ac.uk/cogsys/recognition.html`

tures have the most influence in specific models, for all recognition tasks, as well as reporting correlations between each feature and personality traits in Section 3.2.3.

Task	Classification			Regression			Ranking		
Baseline	n/a	none	50%	n/a	none	0%	n/a	none	50%
Self-report models trained on written data (essays):									
Extraversion	ADA	LIWC	56%	LR	MRC	1%	Rank	LIWC	56%
Emotional stability	SVM	LIWC	58%	M5	LIWC	4%	Rank	LIWC	58%
Agreeableness	SVM	LIWC	56%	LR	LIWC	2%	Rank	LIWC	54%
Conscientiousness	SVM	LIWC	56%	M5	LIWC	2%	Rank	LIWC	56%
Openness to experience	SVM	LIWC	63%	M5	all	7%	Rank	LIWC	61%
Observer report models trained on spoken data (EAR):									
Extraversion	NB	all	73%	REP	LIWC	24%	Rank	prosody	74%
Emotional stability	NB	all	74%	M5	prosody	15%	Rank	MRC	61%
Agreeableness	NB	all	61%	M5R	all*	3%	Rank	all	69%
Conscientiousness	NB	all	68%	M5R	LIWC	18%	Rank	all	67%
Openness to experience	NB	prosody	65%	M5	type*	1%	Rank	LIWC	63%

Table 3.26: Comparison of the best models for each trait, for all three recognition tasks. Each table entry contains the algorithm, the feature set and the model performance. See Sections 3.2.2 and 3.2.4 for details. Depending on the task, the evaluation metric is either the (1) classification accuracy; (2) percentage of improvement over the regression baseline; (3) percentage of correctly ordered pairs (1−ranking loss). Asterisks indicate results that are not significant at the $p < .05$ level.

Although the parameters of the algorithms have not been optimised, the bottom of Table 3.26 seems to indicate that simple models like Naive Bayes or regression trees tend to outperform more complex ones (e.g. Adaboost), confirming results from Oberlander and Nowson [2006]. However, our experiments on the larger essays corpus (more than 2,400 texts) show that support vector machines and boosting algorithms produce higher classification accuracies. It is therefore likely that those algorithms would also perform better on spoken data if they were trained on a much larger corpus than the EAR dataset.

We hypothesised that models of observed personality will outperform models of self-assessed personality. Our results do suggest that observed personality may be easier to model than self-reports, at least in conversational data. For the EAR corpus, we find many good results with models of observed personality, while models of self-assessed personality never outperform the baseline. This may be due to objective observers using the same input as our models (transcripts and audio data), while self-reports are influenced by other factors such as the user's past history and the desirability of the trait [Edwards, 1953].

As there are discrepancies between markers of self-assessed and observed personality, another issue is the identification of the most appropriate model given a specific application. The *true* personality can be approximated by either observer or self-reports, however it is likely that for a specific trait one type of report will be closer to the truth. A hypothesis that remains to be tested is that traits with a high visibility (e.g. extraversion) are more accurately assessed using observer reports, as they tend to yield a higher inter-judge agreement [Funder, 1995], while low visibility traits (e.g. emotional stability) are better assessed by oneself. A personality recogniser aiming to estimate the true personality would therefore have to switch from observer models to self-report models, depending on the trait under assessment.

Beyond practical applications of personality recognition models, this work is also an attempt to explore different ways of looking at the relation between personality and language. We looked at various personality recognition tasks, which vary in complexity: a ranking model can be directly derived from a regression model, while a classification model can be derived from either a ranking or a regression model. Is any type of model closer to the actual relation between language—and more generally behaviour—and personality? Does personality vary continuously, or are there clusters of people with similar trait combinations? If the relation is continuous, classification algorithms will never be able to produce accurate models for more than two classes, because they do not take into account any ordering between the classes. As ranking models outperform classifiers (see Section 3.6), and given the wide range of individual differences reflected by the literature on the Big Five [Allport and Odbert, 1936, Norman, 1963, Goldberg, 1990], we believe that personality varies continuously among members of the population, suggesting that regression or ranking models should be more accurate in the long run. This hypothesis is supported by recent work in medical research showing that antisocial personality disorder varies continuously [Marcus et al., 2006]. Regression provides the most detailed model of the output variables, but depending on whether absolute differences between personality scores are meaningful, or if only relative orderings between people matter, ranking may be more appropriate.

An interesting piece of future work would be to improve these models and examine how well they perform across dialogue domains, which we explore in Chapter 8 (Section 8.4). It is not clear whether the accuracies are high enough to be useful. Applications involving speech recognition will introduce noise in all features except for the prosodic features, probably reducing model accuracy, but since the EAR corpus is relatively small, we expect that more training data would improve performance. Additionally, we believe that the inclusion of gender as a feature would produce better models, as language correlates of perceived personality were shown to depend on the gender of the speaker [Mehl et al., 2006]. Future work should also investigate the combination of individual features in a trait-dependent way. Furthermore, feature selection algorithms could be used to evaluate whether n-gram features are useful, although it is important to perform this selection on a development set to avoid overfitting. Another issue is the poor performance of the utterance type features—since there were significant correlation results for these features in Section 3.2.3, it is unclear why these features are not useful in the statistical models. This could possibly arise from the small size of the dataset, or from the relatively low accuracy of our handcrafted automatic tagger, compared to other work using supervised learning methods [Stolcke et al., 2000, Webb et al., 2005].

Part II

Generating a Recognisable System Personality through Language

Chapter 4

From Personality Markers to Generation Decisions

The first part of this thesis has focused on the recognition of personality through language, i.e. learning a many-to-one mapping between linguistic markers and personality traits. We now turn to the opposite task, i.e. the projection of personality in generated utterances, which requires a one-to-many mapping between the target personality and its linguistic markers in an utterance. One might think that this mapping could be obtained by 'reversing' the models learnt in the previous chapter. However, this approach ignores the first requirement of any natural language generation (NLG) system: to produce grammatical and natural outputs that respect the input meaning representation. While recognition models can detect the personality of ungrammatical utterances, the generation task leaves no room for error, as inconsistencies are likely to be rejected by the user and affect the perception of the intended personality.

A consequence is that a top-down approach is necessary in order to ensure a grammatical *variation space* (i.e. the space of all possible generated outputs given an input meaning representation), e.g. by building a base generator following the standard NLG pipeline [Reiter and Dale, 2000]. As the grammaticality of the output is guaranteed by the base generator, the variation space can then be explored by controlling the generation parameters using various knowledge sources (see Chapters 6 to 9). This view is certainly naive, as inconsistent generation parameters are likely to affect the utterance's grammaticality and naturalness. Nevertheless, we believe that this can be achieved by selecting generation parameters related to the expression of the target variation dimensions—personality traits—that do not affect grammaticality regardless of their value. As far as the naturalness is concerned, it is important to note that a high naturalness cannot be enforced when generating extreme personality traits that are inherently unusual. Therefore, although we aim at producing natural utterances, our primary goal is the exploration of the variation space.

An objective of this thesis is to produce a highly parameterisable generator whose outputs vary along personality dimensions. A first hypothesis is that such language can be generated by varying parameters suggested by psycholinguistic research. This chapter thus *summarises* and *organises* findings from psychological studies about how personality affects language, by mapping them to potential generation parameters within the standard NLG architecture.

4.1 Personality marker studies

The objective of studies assessing how personality affects language is to learn something about human beings, not to produce a computational model of how to generate and recognise language manifesting personality. Most studies are therefore correlational, i.e. they identify correlations between the personality ratings and the linguistic markers of a set of individuals. This chapter is the result of a large organisational effort to compile studies spreading over 50 years, which are listed in Table 4.1 for future reference. These studies typically focus on a single *source of language* and use a specific *personality assessment method*.

4.1.1 Sources of language

Psychologists have studied many forms of language: daily-life conversations, stream-of-consciousness essays, written and oral examinations, emails and even weblogs (see Chapter 3 and Table 4.1). Different types of language provide different means for expressing one's personality. Typically, personality is conveyed more strongly in less constrained forms of language (e.g. daily-life conversations) as genre-specific constraints (e.g. formality) can hide variables that could have marked the speaker's personality. For example, an individual's extravert character is more likely to be de-

Ref	Study	Language	Cues	Assessment	Dimensions
1	Furnham [1990]*	spoken	speech and linguistic markers	self-report	extraversion, type A behaviour, self-monitoring
2	Scherer [1979]*	spoken	speech markers	self-report, emotion induction	extraversion, emotional stability, anxiety *inter alia*
3	Pennebaker and King [1999]	essays	content-analysis category counts	self-report	Big Five traits
4	Dewaele and Furnham [1999]*	spoken	various	self-report	extraversion
5	Oberlander and Gill [2006]	emails	content-analysis category and n-gram counts	self-report	extraversion, neuroticism, psychoticism
6	Mehl et al. [2006]	daily-life conversations	content-analysis category counts	observer, self-report	Big Five traits
7	Siegman and Pope [1965]	spoken	verbal fluency	self-report	extraversion
8	Oberlander and Gill [2004a]	emails	part-of-speech n-grams	self-report	extraversion, neuroticism, psychoticism
9	Oberlander and Gill [2004b]	emails	content-analysis category and n-gram counts	self-report	extraversion, neuroticism
10	Weaver [1998]	questionnaires	communicative behaviour	self-report	extraversion, neuroticism, psychoticism
11	Heylighen and Dewaele [2002]	essays and oral examinations	measure of formality	self-report	extraversion
12	Nowson [2006]	blogs	content-analysis categories and n-grams	self-report	Big Five traits
13	Cope [1969]	spoken	output size, type-token ratio	self-report	extraversion
14	Thorne [1987]	spoken	polarity, focus	self-report	extraversion
15	Siegman [1978]*	spoken	speech markers	various	socio-economic background, extraversion, anxiety, anger, *inter alia*
16	Scherer [1981]*	spoken	speech markers	various	stress, anxiety
17	Gill and Oberlander [2003]	emails	part-of-speech n-gram counts	self-report	extraversion, neuroticism
18	Infante [1995]*	spoken	communicative behaviour	emotion induction	verbal aggressiveness

Table 4.1: Psychological studies on the identification of personality markers in language. An asterisk indicates a review, rather than a specific study.

tected in casual conversations than in a written essay. Despite these limitations, we believe all sources of language can reveal important markers of personality, even in written form [Pennebaker and King, 1999, *inter alia*]. We therefore choose to consider all sources of language for informing our generator, although we aim at projecting personality in dialogue. We thus make the following assumption:

Hypothesis 1 *Personality markers identified in specific forms of language generalise to dialogue utterances.*

Although this hypothesis is not strictly true, it is a useful approximation, as the rarity of personality studies implies that any information about a personality cue is preferable to no information at all. This assumption also gives us access to many significant text-based studies, as the collection of textual data is much easier than dialogue. Regardless of its validity, the usefulness of this approximation for projecting personality in dialogue is tested in Chapter 6 through human evaluation.

4.1.2 Personality assessment methods

There is no gold standard for measuring someone's personality, only different techniques that provide insight into the *true* personality of the subject [Funder, 1997]. These include the following assessment methods (the first two have already been studied in Chapter 3):

- **Self-reports:** The subject assesses his or her own personality by completing a questionnaire [e.g. McCrae and Costa, 1987, John and Srivastava, 1999, Gosling et al., 2003];

- **Observer reports:** Observers assess the personality of the subject by completing a questionnaire;

- **Non-projective tests:** The subject performs a test without knowing how its outcome affects the assessment of personality (e.g. the Rorschach inkblot test);

- **External data:** Personality is inferred from objective information about the subject, such as the subject's employment history;

- **Emotion induction:** Emotions are induced in the subject, and his or her reactions are analysed. This method is only valid for assessing emotions that relate to personality traits.

Because they require less resources, most work on personality markers has used the first two assessment methods, i.e. by studying how linguistic markers correlate with questionnaire ratings of self-reports and observer reports of personality. The induction of emotion is also a common practise for studying markers of short-lived emotions or moods, which can then be related to personality traits. We do not consider other types of studies here.

Over the past decades, the Big Five model has emerged as a standard for evaluating personality [Norman, 1963]. As we already use this framework for personality recognition in Chapter 3, we also use it for controlling the personality in a language generator. The Big Five model consists of the following dimensions (repeated from Chapter 2):

- Extraversion vs. introversion (sociable, assertive vs. reserved, shy)
- Emotional stability vs. neuroticism (calm vs. insecure, anxious)
- Agreeableness vs. disagreeable (friendly vs. antagonistic, faultfinding)
- Conscientiousness vs. unconscientious (organised vs. inefficient, careless)
- Openness to experience (intellectual, insightful vs. shallow, unimaginative)

As the aim of this thesis is to control the user's perception of the system's personality, studies identifying markers of observed personality are the most relevant. However, the Brunswikian lens model [Brunswik, 1956] suggests that cues of self-reported personality tend to align with the cues used by observers over time, in order for those cues to gain usefulness. Additionally, it seems reasonable to assume that personality is conveyed more accurately using cues of self-reported personality rather than using no cues at all. For these reasons, we consider both markers of self-reported and observed personality for building our generator. This decision is based on a second assumption:

Hypothesis 2 *Linguistic markers of self-assessed personality can be perceived by observers.*

This hypothesis will also be tested in Chapter 6, by evaluating whether markers of self-reported personality can be used to convey that personality in a generator.

4.2 NLG parameter mapping

The following sections summarise and organise findings about linguistic markers of each Big Five trait from the studies referenced in Table 4.1, by systematically mapping them to generation decisions in a language generator. When needed, we assume that the goal of the generator is to present information to the user, as it is one of the most common functions of NLG systems. The findings are partitioned according to the following NLG architecture, whose input is a set of high-level communicative goals. Although there exists many different generation frameworks, the one presented here is similar to the standard architecture detailed in Section 2.4.2 in Chapter 2 [Reiter and Dale, 2000]:

- **Content planning:** refine communicative goals, select and structure the propositional content, e.g. by manipulating a rhetorical structure tree [Mann and Thompson, 1988];
- **Sentence planning:**
  - **Syntactic template selection:** decide what syntactic template to select for expressing each proposition, chosen from a handcrafted generation dictionary;
  - **Aggregation:** decide on how to combine the propositions' syntactic representation together to form the utterance, e.g. by selecting specific cue words for expressing the content plan's rhetorical relations (e.g. JUSTIFY, CONTRAST and INFER);
  - **Pragmatic marker insertion:** insert various markers by transforming the utterance's syntactic representation;
  - **Lexical choice:** for each content word, select the final lexical item among a set of synonyms.
- **Realisation:** convert the utterance's syntactic representation into a string.

The sentence planning module is assumed to produce a fully specified syntactic representation, so that the utterance's personality is independent of the realisation phase.[1] Furthermore, the present work strictly focuses on the projection of personality through linguistic cues. We therefore leave the parameterisation of the speech synthesis module as future work.

The association between a finding and a generation parameter represents an *hypothesis* regarding how the finding can be modelled in our domain. As these findings were obtained from various sources of language, it is not clear whether they can be reproduced through a single utterance presenting information to the user. Furthermore, while studies are typically based on surface realisations, generation decisions generally require an interpretation of the cause of these findings, e.g. a high word count can be associated with the expression of a more diverse content and/or more repetitions.

The information presentation task can be seen as communicating the result of a database query to the user. As the goal of this chapter is to present conceptual generation parameters that can be used in different applications, we do not add any further specification to the generation domain. For the same reason, the function of the parameters presented here is only specified at a high level, without any implementation details. Table 4.2 summarises all the personality-related parameters that are mentioned in the rest of this chapter. To fix the ideas, parameters are illustrated through specific examples of evaluative utterances, however each finding can generally be modelled in many other ways. Chapter 5 presents an implementation of each parameter in a concrete information presentation domain, in which the generator provides recommendations for a selection of restaurants. The next sections present the mappings between personality markers derived from the psychology literature and specific generation parameters, for each personality trait in the Big Five framework.

4.3 Extraversion

Extraverts tend to engage in social interaction, they are enthusiastic, risk-taking, talkative and assertive, whereas introverts are more reserved and solitary. Eysenck

[1]The realisation process is thus ignored during the rest of this chapter.

Parameters	Description
Content planning:	
VERBOSITY	Control the number of propositions in the utterance
RESTATEMENTS	Paraphrase an existing proposition, e.g. *'X has great Y, it has fantastic Z'*
REPETITIONS	Repeat an existing proposition
CONTENT POLARITY	Control the polarity of the propositions expressed, i.e. referring to negative or positive attributes
REPETITION POLARITY	Control the polarity of the restated propositions
CONCESSIONS	Emphasise one attribute over another, e.g. *'even if X has great Z, it has bad Y'*
CONCESSION POLARITY	Determine whether positive or negative attributes are emphasised
POLARISATION	Control whether the expressed polarity is neutral or extreme
POSITIVE CONTENT FIRST	Determine whether positive propositions are uttered first
REQUEST CONFIRMATION	Begin the utterance with a confirmation of the request, e.g. *'did you say X?'*
INITIAL REJECTION	Begin the utterance with a rejection, e.g. *'I'm not sure'*
COMPETENCE MITIGATION	Express the speaker's negative appraisal regarding the hearer's request, e.g. *'everybody knows that . . . '*
Syntactic template selection:	
SELF-REFERENCES	Control the number of first person pronouns
SYNTACTIC COMPLEXITY	Control the syntactic complexity (e.g. syntactic embedding)
TEMPLATE POLARITY	Control the template's connotation (positive or negative)
Aggregation:	
PERIOD	Leave two propositions in their own sentences, e.g. *'X has great Y. It has nice Z.'*
RELATIVE CLAUSE	Join propositions with a relative clause, e.g. *'X, which has great Y, has nice Z'*
WITH CUE WORD	Aggregate propositions using *with*, e.g. *'X has great Y, with nice Z'*
CONJUNCTION	Join propositions using a conjunction, or a comma if more than two propositions
MERGE	Merge the subject and verb of two propositions, e.g. *'X has great Y and nice Z'*
ALSO CUE WORD	Join two propositions using *also*, e.g. *'X has great Y, also it has nice Z'*
CONTRAST - CUE WORD	Contrast two propositions using *while, but, however, on the other hand*, e.g. *'While X has great Y, it has bad Z'*, *'X has great Y, but it has bad Z'*
JUSTIFY - CUE WORD	Justify a proposition using *because, since, so*, e.g. *'X is the best, since it has great Y'*
CONCEDE - CUE WORD	Concede a proposition using *although, even if, but/though*, e.g. *'Although X has great Y, it has bad Z'*, *'X has great Y, but it has bad Z though'*
MERGE WITH COMMA	Restate a proposition by repeating only the object, e.g. *'X has great Y, nice Z'*
OBJECT ELLIPSIS	Replace part of a repeated proposition by an ellipsis, e.g. *'X has . . . it has great Y'*
Pragmatic marker insertion:	
SUBJECT IMPLICITNESS	Make the presented object implicit by moving its attribute to the subject, e.g. *'the Y is great'*
NEGATION	Negate a verb by replacing its modifier by its antonym, e.g. *'X doesn't have bad Y'*
SOFTENER HEDGES	Insert syntactic elements (*sort of*, *kind of*, *somewhat*, *quite*, *around*, *rather*, *I think that*, *it seems that*, *it seems to me that*) to mitigate the strength of a proposition, e.g. *'X has kind of great Y'* or *'It seems to me that X has rather great Y'*
EMPHASISER HEDGES	Insert syntactic elements (*really*, *basically*, *actually*, *just*) to strengthen a proposition, e.g. *'X has really great Y'* or *'Basically, X just has great Y'*
ACKNOWLEDGMENTS	Insert an initial back-channel (*yeah*, *right*, *ok*, *I see*, *oh*, *well*), e.g. *'Ok, X has great Y'*
FILLED PAUSES	Insert syntactic elements expressing hesitancy (*I mean*, *err*, *mmhm*, *like*, *you know*), e.g. *'Err... X has, like, great Y'*
EXCLAMATION	Insert an exclamation mark, e.g. *'X has great Y!'*
EXPLETIVES	Insert a swear word, e.g. *'the Y is damn great'*
NEAR EXPLETIVES	Insert a near-swear word, e.g. *'the Y is darn great'*
TAG QUESTION	Insert a tag question, e.g. *'the Y is great, isn't it?'*
STUTTERING	Duplicate parts of a content word, e.g. *'X has gr-gr-great Y'*
IN-GROUP MARKER	Refer to the hearer as a member of the same social group, e.g. *pal*, *mate* and *buddy*
PRONOMINALISATION	Replace references to the object by pronouns
Lexical choice:	
LEXICON FREQUENCY	Control the average frequency of use of each content word (e.g. according to frequency counts from a corpus)
LEXICON WORD LENGTH	Control the average number of letters of each content word
VERB STRENGTH	Control the strength of the verbs, e.g. *'I would suggest'* vs. *'I would recommend'*

Table 4.2: Generation parameters that are hypothesised to affect the utterance's personality. Aggregation parameters are duplicated for each rhetorical relation.

et al. [1985] suggest that this trait is associated with a lack of internal arousal: extraverts are thus seeking additional external stimulation, whereas introverts avoid it. Among all Big Five traits, extraversion is the one that has received the most attention in linguistic studies. There are three reasons for this: (1) the extraversion dimension is often considered as the most 'important' one, as it explains the most variance among the adjective descriptors from which the Big Five factors are derived [see Section 2.1.1 in Chapter 2; Goldberg, 1990, *inter alia*], (2) it is also present in most other personality frameworks—e.g. Eysenck et al.'s PEN model [Psychoticism, Extraversion and Neuroticism; 1985]; and (3) extraversion is often considered as the dimension that has the most influence on language, because it is strongly associated with talkativeness and enthusiasm [Furnham, 1990].

The findings about linguistic markers of extraversion are summarised in Table 4.3, together with one or more associated generation parameters that represent our *hypotheses* about how each finding can be mapped to an NLG system. Most generation parameters are based on study results, however some are derived from hypotheses about how a specific trait affects language (indicated by a single asterisk). The right-most columns (e.g. *Intro* and *Extra*) contain the parameter values for expressing each end of the personality dimension, i.e. either introversion or extraversion. As the correlational studies only provide an indicator of the magnitude and trend of the relation between the personality dimension and the linguistic marker, parameter values are currently restricted to the discrete values *low* and *high*. These fuzzy values are used to fix the reader's intuition, a concrete implementation of continuous and binary parameters is presented in Chapter 5. While some parameters might have more importance than others for conveying a specific trait, Chapter 7 presents a correlational analysis that evaluates their individual contribution. The derivation of each parameter is now presented in more detail, for each component of the NLG architecture presented in Section 4.2.

Content planning: Extraverts are more talkative than introverts [Cope, 1969, Furnham, 1990, Pennebaker and King, 1999, Dewaele and Furnham, 1999, Mehl et al., 2006], we thus hypothesise a VERBOSITY parameter controlling the number of propositions expressed in the utterance. As it is not clear whether extraverts actually produce more content, or are just redundant and wordy, a REPETITIONS

Introvert findings	Extravert findings	Ref	Parameters	Intro	Extra
Content planning:					
Single topic	Many topics, higher verbal output	1,3,4, 6,13	VERBOSITY	low	high
Strict selection	Think out loud	1*	RESTATEMENTS	low	high
			REPETITIONS	low	high
Problem talk, dissatisfaction, negative emotion words	Pleasure talk, agreement, compliment, positive emotion words	3,14	CONTENT POLARITY	low	high
			REPETITION POLARITY	low	high
			CONCESSION POLARITY	low	high
			POSITIVE CONTENT FIRST	low	high
Not sympathetic	Sympathetic, concerned about hearer (but not empathetic)	10	REQUEST CONFIRMATION	low	high
Syntactic template selection:					
Elaborated constructions	Simple constructions	1*	SYNTACTIC COMPLEXITY	high	low
Problem talk	Pleasure talk	3	TEMPLATE POLARITY	low	high
Aggregation:					
Few conjunctions	Many conjunctions	8	CONJUNCTION, BUT, ALSO CUE WORD	low	high
Many unfilled pauses	Few unfilled pauses	2,7	PERIOD	high	low
Many uses of *although*	Few uses of *although*	9	ALTHOUGH CUE WORD	high	low
Formal language	Informal language	1*,11	RELATIVE CLAUSE	high	low
Pragmatic marker insertion:					
Many nouns, adjectives, prepositions (explicit)	Many verbs, adverbs, pronouns (implicit)	11	SUBJECT IMPLICITNESS	low	high
Many negations	Few negations	3	NEGATION	high	low
Many tentative words (e.g. *maybe*, *guess*)	Few tentative words	3	SOFTENER HEDGES:		
			·SORT OF, SOMEWHAT, QUITE, RATHER, I THINK THAT, IT SEEMS THAT, IT SEEMS TO ME THAT	high	low
Formal language	Informal language	1*,11	·KIND OF, LIKE	low	high
			ACKNOWLEDGMENTS:		
			·YEAH	low	high
			·RIGHT, OK, I SEE, WELL	high	low
Few swear words	Many swear words	6	NEAR EXPLETIVES	low	high
Many unfilled pauses	Few unfilled pauses	2,7	FILLED PAUSES:	high	low
			· ERR, I MEAN, MMHM, YOU KNOW		
Realism	Exaggeration (e.g. *really*)	9*	EMPHASISER HEDGES:		
			·REALLY, BASICALLY, ACTUALLY, JUST	low	high
			EXCLAMATION	low	high
Not sympathetic	Sympathetic, concerned about hearer, minimise positive face threat	10	TAG QUESTION	low	high
Few words related to humans	Many words related to humans (e.g. *man*, *pal*)	12	IN-GROUP MARKER	low	high
Lexical choice:					
Rich vocabulary	Poor vocabulary	1*,4	LEXICON FREQUENCY	low	high
Longer words	Shorter words	6	LEXICON WORD LENGTH	high	low
Realism	Exaggeration	*	VERB STRENGTH	low	high

Table 4.3: Summary of language cues for extraversion, as well as the corresponding generation parameters. Asterisks indicate hypotheses, rather than results. Referenced studies are detailed in Table 4.1.

parameter is added to produce an exact repetition of a proposition, and a RESTATEMENTS parameter produces a paraphrased repetition.

Extraverts are more positive; introverts are characterised as engaging in more 'problem talk' and expressions of dissatisfaction [Thorne, 1987, Pennebaker and King, 1999]. This positivity can manifest itself through the choice of information presented to the user, which is controlled by a CONTENT POLARITY parameter. In addition, polarity can also be implied by presenting information subjectively, thus a CONCESSION POLARITY parameter controls whether the positive or the negative content is emphasised, such as in *'even if the food is good, it's expensive'* vs. *'even if it is expensive, the food is good'*. Additional emphasis is conveyed using a REPETITION POLARITY parameter, controlling whether positive or negative information is more likely to be repeated in the utterance.

Carenini and Moore [2000] mention that starting with a positive claim facilitates the hearer's understanding, while finishing with it is more effective if the hearer disagrees. A POSITIVE CONTENT FIRST parameter therefore controls whether positive propositions appear first or last. We hypothesise that extraverts begin their utterances with more positive content, as a consequence of their high enthusiasm.

Weaver [1998] shows that extraverts are more sympathetic to other people—i.e. they show more concern—although this sympathy is not related to empathy, as they are not more inclined to feel other people's feelings. Concern for the user can be expressed in the information presentation domain by emphasising the user's request through an explicit confirmation, which is controlled by a REQUEST CONFIRMATION parameter.

Syntactic template selection: The preference of extraverts for positive content can also be expressed through syntactic template selection, by controlling the connotation of the utterance's syntactic representation using a TEMPLATE POLARITY parameter, e.g. a high parameter value would favour the selection of the claim *'X is the best of its kind'* rather than *'X is not as bad as the others'*. As Furnham [1990] hypothesises that extraverts use simpler syntactic constructions, the template selection is also influenced by a SYNTACTIC COMPLEXITY parameter controlling the template's level of subordination, e.g. the claim *'I am sure you would like X'* is more syntactically complex than *'X is the best'*.

Aggregation: Oberlander and Gill [2004a] show that introverts use more conjunctions in their emails. Thus, an introvert system should combine pieces of information using conjunctions, such as *and*, *but* and *also*. Oberlander and Gill [2004b] also find that introversion is associated with the use of the adverbial clause *although*, which can be expressed by selecting the ALTHOUGH CUE WORD operation when conceding a piece of information over another, as opposed to EVEN IF CUE WORD for example. It has also been found that introverts produce more long unfilled pauses [Siegman and Pope, 1965, Scherer, 1979], which can be controlled at the aggregation level by enforcing that the utterance's propositions are expressed in separate sentences, using the PERIOD aggregation operation. Finally, we hypothesise that introverts produce more relative clauses, as a manifestation of their preference for formal language.

Pragmatic marker insertion: Psychological studies identify many pragmatic markers of extraversion which only affect the utterance locally, and can thus be implemented as separate syntactic transformations. These studies show that introverts produce more negations, tentative words (e.g. *maybe*, *perhaps*) and filled pauses [Pennebaker and King, 1999, Siegman and Pope, 1965, Scherer, 1979]. Negations can be controlled—while preserving the original meaning—by a NEGATION parameter that negates the logical inverse of a proposition, e.g. by producing *'X is not bad'* rather than *'X is good'*. Tentativeness can be expressed through hedging expressions that mitigate the impact of the speaker's statement—referred to as SOFTENER HEDGES—including *sort of*, *somewhat*, *rather*, *I think that*, etc. Filled pauses can be expressed linguistically by inserting the adjuncts *err*, *mmhm*, *I mean*, *like* and *you know*, which are all placed under the FILLED PAUSES category.[2]

Extraverts use more informal, implicit language [Heylighen and Dewaele, 2002]. We associate informal language with the use of adverbial hedges such as *kind of* and *like*, as well as acknowledgments such as *yeah* (as opposed to *well* or *right* for example). Implicitness can be conveyed in the information presentation domain by referring to the object of interest implicitly through its attributes, such as in *'the food is good'* (X is implicit) vs. *'X has good food'* (X is explicit). This syntactic transformation is controlled by the SUBJECT IMPLICITNESS parameter. Oberlander and

[2]As the hedge *you know* can have many other functions, we generally consider it as a filled pause, while we model it individually when needed (e.g. for projecting agreeableness in Section 4.5).

Gill [2004b] also find that extravert emails contain more occurrences of *'I really'* as well as more exclamation marks, suggesting the need for parameters controlling the insertion of adverbs such as *really*, *basically*, *actually* and *just*—referred to as EMPHASISER HEDGES—as well as an EXCLAMATION parameter ending the utterance with an exclamation mark.

Extraversion is also associated with more swearing and references to humans [Oberlander and Gill, 2004b, Mehl et al., 2006, Nowson, 2006]. The use of (near-) swear words can be manipulated by inserting modifiers—e.g. *'the food is darn good'*—given a high NEAR EXPLETIVES parameter value. We use 'near' expletives to avoid conflicting with the positivity associated with extravert language. References to humans can be added locally as adjunct nouns—e.g. *'the food is good pal'*—using an IN-GROUP MARKER parameter. This linguistic marker can also be interpreted as the minimisation of the positive face threat according to Brown and Levinson's politeness theory [1987]. Tag questions also fulfill the same politeness function, as well as contributing to the extravert's expression of sympathy [Weaver, 1998]. They can therefore be inserted automatically using a TAG QUESTION parameter, such as in *'X has good food, doesn't it?'*.

Lexical choice: Introverts use richer and longer words [Furnham, 1990, Dewaele and Furnham, 1999]. These aspects of the speaker's vocabulary can be controlled by a LEXICON FREQUENCY parameter and a LEXICON WORD LENGTH parameter, respectively biasing the selection of content words depending on their frequency of use and their length. Finally, we hypothesise that extraverts produce more exaggerations—as a consequence of their enthusiasm—which results in the use of stronger verbs, e.g. by favouring *love* over *like* in the utterance *'I think you would love X'*.

4.4 Emotional stability

Emotional stability—or neuroticism—is the second most studied personality trait, as it is part of most existing frameworks of personality, such as the Big Five and the PEN model [Norman, 1963, Eysenck et al., 1985]. Neurotics tend to be anxious, negative and oversensitive, whereas emotionally stable people are calm and even-tempered. Eysenck et al. [1985] suggest that this dimension is related to activation thresholds in the nervous system, i.e. neurotics turn more easily into a 'fight-or-flight' state when facing danger, resulting in an increase of their heart beat, muscular tension, level of sweating, etc. In order to increase the number of relevant findings, studies focusing on short-lived emotions that are symptomatic of the personality trait under study are also included, e.g. markers of anxiety are considered as valid markers of neuroticism.[3] This assumption can be summarised as:

Hypothesis 3 *Linguistic markers of emotions that are concomitant of a personality trait can be used to convey that trait in dialogue.*

This hypothesis is motivated by the association of personality traits with a tendency to express or repress specific emotions over a long period [Watson and Clark, 1992].

Content planning: Even more than introversion, neuroticism is largely associated with negativity [Pennebaker and King, 1999], which can thus be controlled by the same polarity parameters—i.e. CONTENT POLARITY, CONCESSION POLARITY and REPETITION POLARITY. See Table 4.4. Neurotics also produce more lexical repetitions [Scherer, 1981], with a lower type-token ratio [Gill and Oberlander, 2003]. Thus, a neurotic generator would have a high REPETITIONS parameter value. Additionally, we hypothesise that their overall lack of control makes neurotics more likely to present a positive claim first in their utterances—i.e. a high POSITIVE CONTENT FIRST parameter value, whereas more stable individuals would finish their utterances with more positive content to have a higher argumentative impact. Following the same assumption, we associate neuroticism with a high POLARISATION parameter value, i.e. the production of more extreme content (regardless of whether it is positive or negative). Finally, we hypothesise that anxiety can be projected in the information presentation domain through explicit requests for confirmation as well as request rejections, producing utterances beginning with *'I'm not sure... did you say X?'* for example. The insertion of these markers is respectively controlled by the REQUEST CONFIRMATION and INITIAL REJECTION parameters.

[3]The term 'anxiety' is sometimes used to describe either an emotion or a permanent trait, the former is then referred to as *state anxiety* and the latter as *trait anxiety*.

Neurotic findings	Stable findings	Ref	Parameters	Neuro	Emot
Content planning:					
Problem talk, dissatisfaction	Pleasure talk, agreement, compliment	3	CONTENT POLARITY	low	high
			REPETITION POLARITY	low	high
			CONCESSION POLARITY	low	high
Direct claim	Inferred claim	*	POSITIVE CONTENT FIRST	high	low
High verbal productivity	Low verbal productivity	15	VERBOSITY	high	low
Many lexical repetitions	Few lexical repetitions	9,16	REPETITIONS	high	low
Polarised content	Neutral content	*	POLARISATION	high	low
Stressed	Calm	*	REQUEST CONFIRMATION	low	high
			INITIAL REJECTION	high	low
Syntactic template selection:					
Many self-references	Few self-references	3,6,9	SELF-REFERENCES	high	low
Problem talk	Pleasure talk	3	TEMPLATE POLARITY	low	high
Aggregation:					
Low use of 'punct *which*'	High use of 'punct *which*'	9	RELATIVE CLAUSE	low	high
Many conjunctions	Few conjunctions	8	MERGE	high	low
Few short silent pauses	Many short silent pauses	15	CONJUNCTION	low	high
Low use of 'punct *so*'	High use of 'punct *so*'	9	JUSTIFY - SO CUE WORD	low	high
Low use of clause final *also*	High use of clause final *also*	9	INFER - ALSO CUE WORD	low	high
Many inclusive words (e.g. *with*, *and*)	Few inclusive words	9,17	WITH CUE WORD	high	low
High use of final *though*	Low use of final *though*	8	CONCEDE - BUT/THOUGH CUE WORD	high	low
Many long silent pauses	Few long silent pauses	15	PERIOD	high	low
Many 'non-ah' disfluencies (omission)	Few 'non-ah' disfluencies	16**	RESTATE - OBJECT ELLIPSIS	high	low
Pragmatic marker insertion:					
Many pronouns, few articles	Few pronouns, many articles	3,8	SUBJECT IMPLICITNESS	low	high
			PRONOMINALISATION	high	low
			SOFTENER HEDGES:		
Few tentative words	Many tentative words	6	·SORT OF, SOMEWHAT, QUITE, RATHER, IT SEEMS THAT,IT SEEMS TO ME THAT, KIND OF	low	high
Many self-reference	Few self-references	3,6,9	·I THINK THAT	high	low
Many filled pauses (apprehensive)	Few filled pauses	2,10	FILLED PAUSES: · ERR, I MEAN, MMHM, LIKE	high	low
			ACKNOWLEDGMENTS:		
More aquiescence	Few aquiescence	10	·YEAH, RIGHT, OK	high	low
Many self references	Few self references	3,6,9	·I SEE	high	low
High use of 'punct *well*'	Low use of 'punct *well*'	9	·WELL	high	low
Exaggeration	Realism	*	EMPHASISER HEDGES:		
			·REALLY, ACTUALLY,	high	low
			·BASICALLY, JUST	low	high
Many rhetorical interrogatives	Few rhetorical interrogatives	*	TAG QUESTION	high	low
Frustration	Less frustration	10	EXPLETIVES	high	low
Many 'non-ah' disfluencies (stutter)	Few 'non-ah' disfluencies	16**	STUTTERING	high	low
Lexical choice:					
Many frequent words	Few frequent words	9,17	LEXICON FREQUENCY	high	low
High concreteness	Low concreteness	9,17			
Exaggeration	Realism	*	VERB STRENGTH	high	low

Table 4.4: Summary of language cues for emotional stability, as well as the corresponding generation parameters. One asterisk indicates an hypothesis, rather than a result. Two asterisks indicate a marker of a related emotion (e.g. anxiety). Aggregation parameter names are prefixed with the rhetorical relation they realise.

Syntactic template selection: Studies consistently show that neurotics produce more self-references [Pennebaker and King, 1999, Oberlander and Gill, 2004b, Mehl et al., 2006]. Thus, neuroticism can be conveyed using a SELF-REFERENCES parameter that biases the template selection process by favouring templates with first-person pronouns, such as in the template *'I am sure you would like X'*. Furthermore, as with extraversion, polarity can also be expressed through template selection. Neuroticism can thus be projected by selecting negatively-connotated templates, with a low TEMPLATE POLARITY parameter value.

Aggregation: Emotionally stable people were shown to produce more *which* pronouns in their emails, whereas neurotics prefer the conjunction *and* [Oberlander and Gill, 2004b,a]. The former preference can be modelled by the RELATIVE CLAUSE aggregation operation, such as in the utterance *'X, which has good food, has nice service'*. The production of conjunctions can be controlled by the MERGE operation, which combines propositions together by grouping their objects with a conjunction, e.g. *'X has good food and nice service'*. Siegman [1978] reports that emotionally stable speakers produce more short unfilled pauses, whereas anxious speakers produce longer pauses. Interestingly, these speech cues can be controlled at the aggregation level: short pauses can be realised textually by separating propositions with commas using the CONJUNCTION aggregation operation, while long pauses are conveyed by leaving propositions in separate sentences using the PERIOD operation. Oberlander and Gill [2004a] show that neurotics avoid using the *so* and *also* cue words, while they produce more inclusive words—e.g. *with*, *and*—as well as more occurrences of the adverb *though*. These specific cues can be used for combining, justifying and conceding information using the corresponding aggregation parameters (e.g. CONCEDE - BUT/THOUGH CUE WORD).[4] Finally, Scherer [1981] reports that neurotics are more likely to omit words in their speech. Such disfluencies can be reproduced in the information presentation domain by partially repeating a proposition with ellipsis dots, e.g. *'X has . . . it has good food'*. This linguistic behaviour is controlled at the aggregation level by repeating content using the OBJECT ELLIPSIS aggregation operation.

[4]Aggregation parameter names are prefixed with the rhetorical relation they realise.

Pragmatic marker insertion: As emotionally stable people were found to produce more articles (i.e. common nouns) and fewer pronouns [Oberlander and Gill, 2004a], emotional stability is associated with a high SUBJECT IMPLICITNESS parameter value, and a low PRONOMINALISATION value. Interestingly, emotional stability is also linked to the avoidance of tentative words (e.g. *maybe*, *rather*, *kind of*) [Mehl et al., 2006], which are controlled by the SOFTENER HEDGES parameters. This suggests that neurotics tend not to express their anxiety by verbalising their uncertainties. However, their preference for self-references [Pennebaker and King, 1999] suggests a high use of subordination hedges, such as *I think that*. The literature also shows that neurotics produce more filled pauses [Scherer, 1979, Weaver, 1998], thus a neurotic generator requires high values for the FILLED PAUSES parameters (e.g. *err*, *mmhm*, *I mean*). Scherer [1981] reports that anxiety is also associated with 'non-ah' disfluencies, i.e. alterations of the intended lexical and phrasal output. While word omissions are modelled at the aggregation level, we control whether syllables are repeated within words by adding a STUTTERING parameter. Weaver [1998] shows that neuroticism is associated with frustration and aquiescence, which can be modelled respectively with high EXPLETIVES and ACKNOWLEDGMENTS parameter values. This last finding is confirmed by Oberlander and Gill [2004b], who found that neurotics use the sentence-initial *well* more often. Furthermore, we hypothesise that neurotics are more likely to exaggerate when presenting information, they are thus associated with high EMPHASISER HEDGES parameter values (e.g. *really*, *actually*). Finally, we assume that neurotics use rhetorical questions to reduce their anxiety, which we model with a high TAG QUESTION parameter value.

Lexical choice: As far as their vocabulary is concerned, neurotics use more frequent and concrete words in their emails [Gill and Oberlander, 2003, Oberlander and Gill, 2004b], thus yielding a higher LEXICON FREQUENCY parameter value. As with extraversion, we do not know of any study focusing on the strength of the vocabulary used, we therefore associate neuroticism with the use of stronger verbs—i.e. a higher VERB STRENGTH parameter value—based on the exaggeration hypothesis made in the previous paragraph.

4.5 Agreeableness

Agreeable people are generous, optimistic, emphatic, interested in others, and they make people feel comfortable. On the other hand of the scale, disagreeable people are self-interested, and they do not see others positively. Agreeableness has not been studied as much as extraversion and emotional stability, as it has only emerged with the Big Five framework.

Disagreeable findings	Agreeable findings	Ref	Parameters	Disag	Agree
Content planning:					
Problem talk, dissatisfaction	Pleasure talk, agreement, compliment	3,6	CONTENT POLARITY	low	high
			REPETITION POLARITY	low	high
			CONCESSION POLARITY	low	high
			POSITIVE CONTENT FIRST	low	high
Fewer empathy	More empathy	*	REQUEST CONFIRMATION	low	high
Many personal attacks (competence)	Few personal attacks	18	COMPETENCE MITIGATION	high	low
Many commands, global rejections	Few commands, global rejections	18	INITIAL REJECTION	high	low
Syntactic template selection:					
Problem talk	Pleasure talk	3,6	TEMPLATE POLARITY	low	high
Few self-references	Many self-references	3,6	SELF-REFERENCES	high	low
Aggregation:					
Many pauses	Few pauses	15	PERIOD	high	low
Pragmatic marker insertion:					
Many articles	Few articles	3,6	SUBJECT IMPLICITNESS	high	low
Many negations	Few negations	6	NEGATION	high	low
Many swear words	Few swear words	6,18	EXPLETIVES	high	low
No politeness form	Minimise negative face threat	*	SOFTENER HEDGES: ·SORT OF, SOMEWHAT, QUITE, RATHER, IT SEEMS THAT, IT SEEMS TO ME THAT, AROUND, KIND OF	low	high
Few insight words	Many insight words (e.g. *see*, *think*)	6	·I THINK THAT	low	high
No politeness form	Minimise positive face threat	*	ACKNOWLEDGMENTS: ·YEAH, RIGHT, OK, WELL	low	high
Few insight words	Many insight words	6	·I SEE	low	high
No politeness form	Minimise negative face threat	*	EMPHASISER HEDGES: ·REALLY, BASICALLY, ACTUALLY, JUST	high	low
			FILLED PAUSES:		
No politeness form	Minimise positive face threat	*	·YOU KNOW	low	high
			TAG QUESTION	low	high
			IN-GROUP MARKER	low	high
Lexical choice:					
Few frequent words	Many frequent words	12	LEXICON FREQUENCY	low	high
Shorter words	Longer words	6,12	LEXICON WORD LENGTH	low	high

Table 4.5: Summary of language cues for agreeableness, as well as the corresponding generation parameters. An asterisk indicates an hypothesis, rather than a result.

Content planning: Agreeable people produce more positive emotion words in both essays and conversations [Pennebaker and King, 1999, Mehl et al., 2006], which we model at the content level with high CONTENT POLARITY, REPETITION POLARITY and CONCESSION POLARITY parameter values in Table 4.5. In the information presentation domain, their optimism is associated with the presentation of positive information first (POSITIVE CONTENT FIRST parameter), while their empathy is conveyed by asking for explicit confirmations (REQUEST CONFIRMATION parameter). Infante [1995] reports that verbal aggressiveness is usually expressed through personal attacks, we thus associate disagreeableness with the mitigation of the user's competence by presenting his or her request as trivial, e.g. by making the system's answer begin with *'everybody knows that ...'* (COMPETENCE MITIGATION parameter). The same author shows that disagreeable speakers produce more rejections, which we model with a high INITIAL REJECTION parameter value.

Syntactic template selection: As with other traits, polarity is also modelled using the TEMPLATE POLARITY parameter, to bias the selection of syntactic templates towards positively or negatively-connotated templates. As Mehl et al. [2006] show that self-references are perceived as disagreeable, disagreeableness is associated with a high SELF-REFERENCES parameter value.

Aggregation: Siegman [1978] reports that speakers perceived as warm tend to avoid long unfilled pauses. We therefore associate disagreeableness with utterances containing multiple short sentences, which can be obtained with a high PERIOD aggregation parameter value.

Pragmatic marker insertion: Disagreeable people produce more negations and articles [Pennebaker and King, 1999, Mehl et al., 2006], which we model using the NEGATION and SUBJECT IMPLICITNESS parameters. Mehl et al. also show that both self-reports and observer reports of agreeableness correlate negatively with the use of swear words, our disagreeable generator is thus set to a high EXPLETIVES parameter value. We make a distinction between expletives (e.g. *damn*, *God*) and less harmful 'near-expletives' that are used to convey extraversion (e.g. *darn*, *gosh*). As Pennebaker and King find that agreeable people produce more insight words (e.g.

think, *see*), the subordination hedge I THINK THAT as well as the acknowledgment I SEE are set to high values for that trait.

We hypothesise that agreeableness is also associated with the minimisation of negative face threat according to Brown and Levinson's politeness theory [1987], i.e. agreeable speakers avoid face-threatening acts that reduce the hearer's freedom of action. As a result, agreeable speakers are modelled as producing more understatement and hedges (such as *kind of* or *rather*), with high values for the SOFTENER HEDGES parameters and low values for the EMPHASISER HEDGES parameters. Furthermore, agreeable speakers are hypothesised to minimise the hearer's positive face threat—i.e. by making sure the needs of the hearer are perceived as worthy—by asserting common ground using confirmations and the hedging expression *you know*, as well as using in-group markers and tag questions [see Brown and Levinson, 1987]. Each of these politeness markers is controlled by a specific generation parameter in Table 4.5.

Lexical choice: As agreeable speakers use longer, more frequent words [Mehl et al., 2006, Nowson, 2006], they are modelled with high LEXICON WORD LENGTH and LEXICON FREQUENCY parameter values.

4.6 Conscientiousness

Conscientiousness is related to the control of one's impulses, resulting in careful, self-disciplined, and success-driven people on the one side, and impulsive, disorganised, and laid-back individuals on the other. Similarly to agreeableness, recent work has studied linguistic correlates of conscientiousness, however it has not been researched as extensively as extraversion.

Content planning: Even if polarity is not as strongly related to conscientiousness as to other dimensions, it still plays an important role for projecting that trait. Pennebaker and King [1999] find that conscientious students produce more positive emotion words in their stream-of-consciousness essays, while Mehl et al. [2006] show that negative emotion words are perceived as unconscientious. Thus again, polarity-related parameters are used to convey this difference at the content level. Additionally, we hypothesise conscientiousness (i.e. carefulness) to be associated

Unconscientious findings	Conscientious findings	Ref	Parameters	Unc	Consc
Content planning:					
Few positive emotion words, many negative emotion words	Many positive emotion words (e.g. *happy*, *good*), few negative emotion words (e.g. *hate*, *bad*)	3 6	CONTENT POLARITY REPETITION POLARITY CONCESSION POLARITY	low low low	high high high
Less perspective	More perspective	*	CONCESSIONS	low	high
Less careful	Check that information is conveyed correctly	*	REQUEST CONFIRMATION	low	high
More vague	Straight to the point	*	RESTATEMENTS REPETITIONS INITIAL REJECTION	high high high	low low low
Syntactic template selection:					
Few positive affect	Some positive affect	3,6	TEMPLATE POLARITY	low	high
Aggregation:					
Many exclusive words (e.g. *but*, *without*)	Few exclusive words	3	CONTRAST - ANY CUE WORD	high	low
Many causation words (e.g. *because*, *hence*)	Few causation words	3	JUSTIFY - ANY CUE WORD	high	low
Informal	Formal	*	ALTHOUGH, WHILE, SINCE, HOWEVER CUE WORD RELATIVE CLAUSE	low low	high high
Pragmatic marker insertion:					
Many negations	Few negations	3	NEGATION	high	low
Many swear words	Few swear words	6	EXPLETIVES NEAR EXPLETIVES	high high	low low
Many references to friends (e.g. *pal*, *buddy*)	Few references to friends	12	IN-GROUP MARKER	high	low
Many disfluencies, filler words	Few disfluencies, filler words	6	FILLED PAUSES: · ERR, I MEAN, MMHM, LIKE	high	low
			SOFTENER HEDGES:		
Few insight words	Many insight words	6	· I THINK THAT	low	high
Informal	Formal	*	·KIND OF ·SOMEWHAT, RATHER, SORT OF, QUITE ACKNOWLEDGMENTS: ·I SEE, WELL ·YEAH, KIND OF	high low low high	low high high low
Impulsive	Not impulsive	*	EXCLAMATION	high	low
Lexical choice:					
Many frequent words	Few frequent words	*	LEXICON FREQUENCY	high	low
Shorter words	Longer words	6	LEXICON WORD LENGTH	low	high

Table 4.6: Summary of language cues for conscientiousness, as well as the corresponding generation parameters. Asterisks indicate hypotheses, rather than results.

with explicit confirmations of the user's request, resulting in a high REQUEST CONFIRMATION parameter value in Table 4.6. A second hypothesis is that unconscientious speakers do not bother putting information into perspective, yielding a low CONCESSIONS parameter value. Finally, we assume that conscientious speakers are straight to the point, thus avoiding rejections as well as repetitions of information, i.e. with lower INITIAL REJECTION, REPETITIONS and RESTATEMENTS parameter values.

Syntactic template selection: As conscientiousness has been linked to positive affect [Pennebaker and King, 1999, Mehl et al., 2006, Watson and Clark, 1992], the TEMPLATE POLARITY parameter is used to select templates with a more positive connotation.

Aggregation: Pennebaker and King [1999] show that unconscientious students produce more words related to exclusion (e.g. *but*, *without*) and causation (e.g. *because*, *hence*). Thus, parameters favouring the use of specific cue words for contrasting and supporting information are set to high values for that trait, e.g. CONTRAST - BUT CUE WORD. Because of their thoroughness, we also hypothesise that conscientious speakers use a more formal language, thus producing more formal discourse connectives, e.g. relative clauses and cue words such as *although*, *while*, *however* and *since*.

Pragmatic marker insertion: Unconscientious speakers produce more swear words and negations [Pennebaker and King, 1999, Mehl et al., 2006], they are thus modelled with high EXPLETIVES and NEGATION parameter values. Mehl et al. also find that they produce more disfluencies and filled pauses, suggesting high FILLED PAUSES parameter values (e.g. *err*), but fewer words related to insight—i.e. a low value for the I THINK THAT hedging parameter. Nowson [2006] also shows that unconscientious bloggers produce more references to friends in their posts, which we generalise to dialogue using the IN-GROUP MARKER parameter.

The impulsiveness of unconscientious speakers is hypothesised to express itself through the use of exclamations, although it is not confirmed by any study. Nevertheless, we set our unconscientious generator to a high EXCLAMATION parameter value. Finally, based on the formality assumption made at the aggregation level, we associate conscientiousness with the use of formal softener hedges and acknowledgment markers (e.g. *somewhat*, *right*).

Lexical choice: As Mehl et al. [2006] show that longer words are perceived as more conscientious, the LEXICON WORD LENGTH parameter is set to high for that trait. Furthermore, we make the assumption that conscientious speakers use more formal, infrequent words, yielding a low LEXICON FREQUENCY parameter value.

4.7 Openness to experience

Openness to experience is the trait generating the most controversy in the personality psychology community. It is the weakest factor among the Big Five traits, and it has also been referred to as *intellect* or *culture* in some studies [Peabody and Goldberg, 1989]. People that are open to experience are usually creative, sophisticated, intellectual, curious and interested in art. A lack of openness is thus associated with narrow-mindedness, unimaginativeness and ignorance. As with previous traits, we include findings concerning dimensions that are related to openness to experience/intellect, such as socio-economic background.

Content planning: Although the relation between polarity and openness to experience is not as clear as with other traits, Nowson [2006] shows that open bloggers use more positive emotion words. We thus model open speakers with high values for all polarity parameters in Table 4.7. Watson and Clark [1992] report that openness is generally associated with affect, which we model with a high POLARISATION parameter value to generate more polarised content. As speakers from a higher socio-economic background produce more elaborations [Siegman, 1978], we associate openness to experience with a higher VERBOSITY and a lower INITIAL REJECTION parameter value, based on the assumption that culture and intellect correlate with socio-economic status. Based on the definition of openness to experience, we also hypothesise that open speakers put information into perspective, which is interpreted in our domain as the production of concessions (i.e. a high CONCESSIONS parameter value). Finally, we also assume that they use more politeness forms, resulting in more explicit request confirmations at the content level in Table 4.7.

Syntactic template selection: As with other traits, additional control of the polarity is provided through the TEMPLATE POLARITY parameter. Pennebaker and King [1999] show that open speakers produce fewer self-references, which we model with a low SELF-REFERENCES parameter value. Finally, we associate the intellectual facet of openness to experience with the production of more syntactically complex utterances, resulting in a high SYNTACTIC COMPLEXITY parameter value.

Non-open findings	Open findings	Ref	Parameters	Non-op	Open
Content planning:					
Few positive emotion words	Many positive emotion words (e.g. *happy*, *good*)	12	CONTENT POLARITY	low	high
			REPETITION POLARITY	low	high
			CONCESSION POLARITY	low	high
Low meaning elaboration	High meaning elaboration	15*,**	VERBOSITY	low	high
			INITIAL REJECTION	high	low
Less perspective	More perspective	*	CONCESSIONS	low	high
Few politeness forms	Many politeness forms	*	REQUEST CONFIRMATION	low	high
Syntactic template selection:					
Few positive emotion words	Many positive emotion words	12	TEMPLATE POLARITY	low	high
Many self-references	Few self-references	3	SELF-REFERENCES	high	low
Simple construction	Complex constructions	*	SYNTACTIC COMPLEXITY	low	high
Aggregation:					
Few exclusive words	Many exclusive words (e.g. *but*, *without*)	3	CONTRAST - ANY CUE WORD	low	high
Many causation words (e.g. *because*, *hence*)	Few causation words	3	JUSTIFY - ANY CUE WORD	high	low
Few inclusive words	Many inclusive words (e.g. *with*, *and*)	12	WITH CUE WORD	low	high
			CONJUNCTION	low	high
			MERGE	low	high
Simple construction	Complex constructions	*	RELATIVE CLAUSE	low	high
Many planning errors	Few planning errors	*	RESTATE - OBJECT ELLIPSIS	high	low
Pragmatic marker insertion:					
Few articles, many third person pronouns	Many articles, few third person pronouns	3,6	SUBJECT IMPLICITNESS	low	high
			PRONOMINALISATION	high	low
Few tentative words	Many tentative words (e.g. *maybe*, *guess*)	3,6	SOFTENER HEDGES:		
			·SORT OF, SOMEWHAT, QUITE, RATHER, IT SEEMS THAT, IT SEEMS TO ME THAT, AROUND, KIND OF	low	high
Few insight words	Many insight words (e.g. *think*, *see*)	3,6	· I THINK THAT	low	high
			ACKNOWLEDGMENTS:		
			·I SEE	low	high
Many filler words and within-utterance pauses	Few filler words and within-utterance pauses	*,15**	FILLED PAUSES:	high	low
			·ERR, I MEAN, MMHM, LIKE		
Few politeness forms	Many politeness forms	*	TAG QUESTION	low	high
			NEAR EXPLETIVES	high	low
Lexical choice:					
More frequent words, lower age of acquisition	Less frequent words, higher age of acquisition	*,12	LEXICON FREQUENCY	high	low
Shorter words	Longer words	3,6,12	LEXICON WORD LENGTH	low	high
Milder verbs	Stronger, uncommon verbs	*	VERB STRENGTH	low	high

Table 4.7: Summary of language cues for openness to experience, as well as the corresponding generation parameters. One asterisk indicates an hypothesis, rather than a result. Two asterisks indicate a marker of a facet associated with that trait (e.g. socio-economic background).

Aggregation: Pennebaker and King [1999] show that open speakers use more words related to exclusion (e.g. *but*, *without*), but fewer words related to causation (e.g. *because*, *so*). These findings are modelled by controlling whether specific cue words are used for contrasting and supporting pieces of information (i.e. with CONTRAST and JUSTIFY rhetorical relations), rather than presenting each information in a separate sentence (with the PERIOD aggregation parameter). Nowson [2006] also find that open bloggers produce more inclusive words (e.g. *with*, *and*), which are controlled using the WITH CUE WORD, CONJUNCTION and MERGE aggregation operations. As at the template selection level, we hypothesise that open speakers produce more complex syntactic structures through subordination, resulting in a high RELATIVE CLAUSE parameter. Similarly, we assume that speakers low on openness to experience produce more omission disfluencies, which we model using the OBJECT ELLIPSIS aggregation operation when repeating information. For example, this parameter generates repetitions such as *'X has . . . it has good food'*.

Pragmatic marker insertion: Open speakers produce more articles but fewer third person pronouns [Pennebaker and King, 1999, Mehl et al., 2006], thus our open generator generates implicit references with a high SUBJECT IMPLICITNESS parameter value—e.g. producing *'the food is good'* as opposed to *'it has good food'*. Whether or not references to the entity under focus are pronominalised is controlled by the PRONOMINALISATION parameter, which is thus set to low to convey openness. The same authors find that openness to experience is associated with words related to tentativeness (e.g. *maybe*, *perhaps*) and insight (e.g. *think*, *see*), which we control using softener hedges such as *kind of* and *I think that*, and the acknowledgment *I see*. As in the previous paragraph, speakers low on openness to experience are hypothesised to produce more disfluencies, which are modelled by the insertion of filler words using high FILLED PAUSES parameter values. This assumption is strengthened by studies reported by Siegman [1978], showing that a low vocabulary proficiency is associated with many pauses within the utterance. We also assume that open speakers are more polite—based on the cultural facet of the trait—which results in a high TAG QUESTION and a low NEAR EXPLETIVES parameter value for that end of the scale.

Lexical choice: Openness to experience is consistently associated with the use of longer words, for both self-reported and observed perceptions of written and spoken language [Pennebaker and King, 1999, Mehl et al., 2006, Nowson, 2006]. Our open generator is thus set to a high LEXICON WORD LENGTH parameter value. Nowson also finds that open bloggers use words that are acquired later in childhood, which we associate with a lower frequency of use (i.e. a low LEXICON FREQUENCY parameter value). This parameter setting is also motivated by the association of openness to experience with a richer vocabulary in some questionnaires [McCrae and Costa, 1987]. Finally, we hypothesise that the sophistication facet of openness to experience results in the use of stronger verbs in evaluative utterances—such as *love* in *'I am sure you would love X'*—which we model using the VERB STRENGTH parameter.

4.8 Summary

This chapter presents and organises findings from psychological studies in order to explore their use for conveying personality in a natural language generation system. We systematically map vague and genre-specific findings to generation decisions that can be implemented in a concrete generator. A consequence is that each mapping represents a *hypothesis* about how the finding can be modelled in a computational framework and *generalise* to a specific application domain. Some traits—e.g. extraversion—were more widely studied than others, their mappings are thus more comprehensive and more likely to generalise. To address this issue, a set of hypotheses were made in order to maximise the number of findings related to each trait, thus increasing the potential linguistic variation. Section 7.2.1 in Chapter 7 presents a correlational analysis that evaluates whether individual generation decisions successfully convey the intended personality.

This chapter presents a principled approach for deriving high-level generation parameters that can project personality in any information presentation domain. This methodology is an essential aspect of this work, as the implementation, control, and effect of each parameter will be investigated throughout the rest of this thesis. The next chapter presents a concrete implementation of all the parameters in a specific domain, thus providing a testable computational model for evaluating whether the parameter mappings and their underlying hypotheses can be used to project personality in dialogue. Such an evaluation is presented in Chapter 6, while Chapters 8 and 9 present more advanced statistical methods for controlling the generated personality.

Chapter 5

Implementing Personality Markers in a Natural Language Generator

The previous chapter has reviewed markers of personality identified in the psychology literature, and associated them with generation decisions. While this mapping gives insight into what a language generator should do to express personality, it does not specify how to do so. The current chapter addresses this issue by focusing on how these generation decisions are implemented in the PERSONAGE language generator.

5.1 Framework overview

Our method for generating personality consists of two main components: (1) the PERSONAGE base generator, which produces language expressing various personality traits by implementing the generation decisions described in Chapter 4, and (2) a *personality model* of how the generation decisions influence the projected personality.

The following chapters evaluate three alternative approaches to personality generation, each with a different personality model controlling the base generator. The *rule-based* approach associates each personality type with a set of abstract parameter values derived from psychological studies,[1] and the setting associated

[1] Parameter values are abstract in the sense that they only indicate a trend as opposed to exact

with the desired trait is selected at generation time. This personality model can therefore target each end of the Big Five scales. The derivation of the parameter sets is detailed in Chapter 4, and an evaluation of the rule-based approach is reported in Chapter 6. The second method, referred to as *overgenerate and select*, requires the generation of many candidate utterances using the base generator by randomly varying its input parameters. The personality model is then used to predict the personality score of each utterance, and the utterance yielding the closest score to the target is selected. An evaluation of this method can be found in Chapter 8. In the third approach, *parameter estimation models* estimate the optimal generation parameters given target personality scores, which are then used by the base generator to produce the output utterance. This method is presented and evaluated in Chapter 9. The last two approaches model personality variation continuously, whereas the rule-based method only produces extreme personality. The overall methodology behind PERSONAGE can be summarised in the following steps:

1. Developing the base generator:
   (a) Identify personality markers from psychological studies (see Chapter 4);
   (b) Map these markers to natural language generation decisions;
   - **Rule-based generation mode:** derive parameter settings for both ends of each trait from the mappings defined in (b), and use them for generating extreme personality (see Chapters 4 and 6).
2. Training the generator's personality model:
   (a) Generate utterances covering the full parameter range (see Chapter 7);
   (b) Judges rate the output of step (a) with a standard personality test;
   (c) Compute feature values for each utterance based on the actual decisions of the generator and possibly other utterance features (see Chapter 7);
   - **Overgenerate and select mode:** train a statistical model to predict the judges' ratings from the features, and use it to rank randomly generated utterances based on the target personality (see Chapter 8).

generation decisions, because of the imprecise nature of the psychology findings.

- **Parameter estimation mode:** train a statistical model to predict the generation decisions from the judges' ratings, and generate the target personality using the predicted parameter values (see Chapter 9).

While Chapters 6, 8 and 9 detail the personality models and their application, the present chapter focuses on the development of the PERSONAGE base generator. The following sections thus describe the generator's domain, overall architecture and implementation.

5.2 Projecting personality in a specific domain

Analysing someone's personality accurately requires having access to *good information* about his or her behaviour [Funder, 1995]. If one's observable behaviour is restricted to conversation turns, it seems reasonable to assume that some types of conversation—or dialogue system *domains*—are more likely to reveal the speaker's personality than others. For example, an argumentative utterance might be more informative about the speaker's personality than an answer to a factoid question. However, even the latter can convey pragmatic variation, such as in *'Err... I would think that the answer is 42, I guess'* or *'Yeah that's 42!'*.

PERSONAGE builds on some aspects of the SPARKY sentence planner [Stent et al., 2004], which provides comparisons and recommendations of restaurants in New York City. While keeping the same domain in our experiments, the generation parameters are designed to generalise to any evaluative utterance in the *information presentation domain*. We hypothesise that evaluative utterances are suitable for expressing recognisable personality, as they allow for substantial variation of the utterance length, polarity and subjectivity of the opinion expressed, all of which were shown to correlate with various personality traits in Chapter 4.

5.3 Input structure

PERSONAGE's input consists of a selection of restaurants in New York City, with associated scalar values representing evaluative ratings for six attributes: *food quality*, *service*, *cuisine*, *location*, *price* and *atmosphere*.[2]

[2]The attribute values used in the present work are derived from Zagat Survey's ratings, and mapped from a 30-point scale to the $[0, 1]$ interval.

The second input consists of parameter values for the generation parameters defined in Chapter 4. As part of the purpose of an independent language generator is to be re-usable in different applications, an objective of the current work is to make PERSONAGE as domain-independent as possible. A consequence is that parameter values are normalised between 0 and 1 for continuous parameters, and to 0 or 1 for binary parameters. For example, a VERBOSITY parameter set to 1 maximises the utterance's verbosity given the input, regardless of the actual number of propositions expressed.

5.4 PERSONAGE's architecture

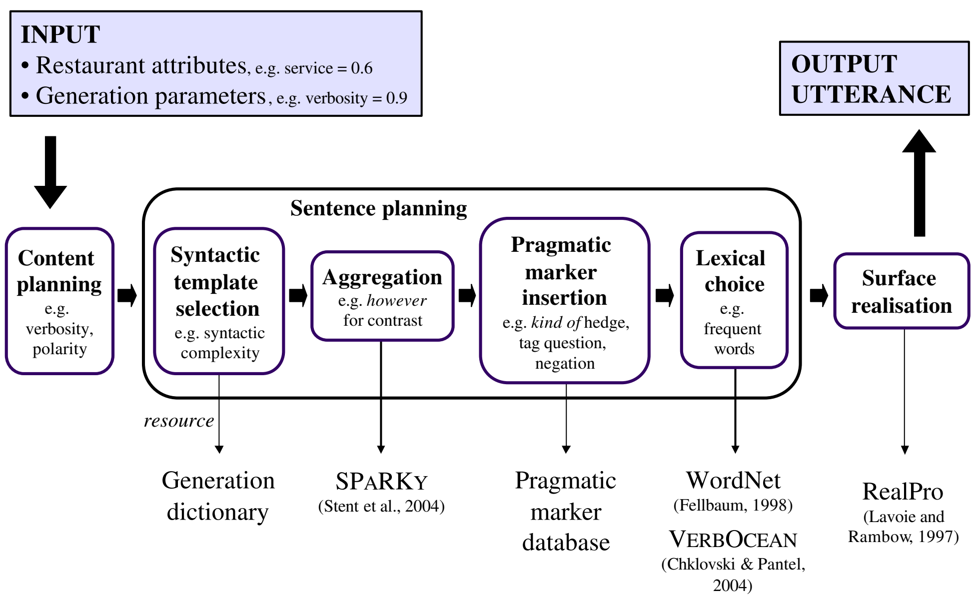


Figure 5.1: The architecture of the PERSONAGE base generator.

PERSONAGE implements the traditional pipelined natural language generation (NLG) architecture [Reiter and Dale, 2000], which consists of a series of sequential components, with each component processing the output of its predecessor. The high-level architecture of the base generator—outlined in Section 4.2 in Chapter 4—is illustrated in Figure 5.1. The first component is the *content planner*,[3] which specifies the structure of the information to be conveyed. The resulting content plan tree is then processed by the *sentence planner*,[4] which selects syntactic

[3]The content planning phase is sometimes referred to as *text planning*.
[4]The sentence planning phase is sometimes referred to as *micro-planning*.

templates for expressing individual propositions, and aggregates them to produce the utterance's full syntactic structure. The pragmatic marker insertion component then modifies the syntactic structure locally to produce various pragmatic effects, depending on the markers' insertion constraints. The next component selects the most appropriate lexeme for each content word, given the lexical selection parameters. Finally, the RealPro realiser [Lavoie and Rambow, 1997] converts the final syntactic structure into a string by applying surface grammatical rules, such as morphological inflection and function word insertion.

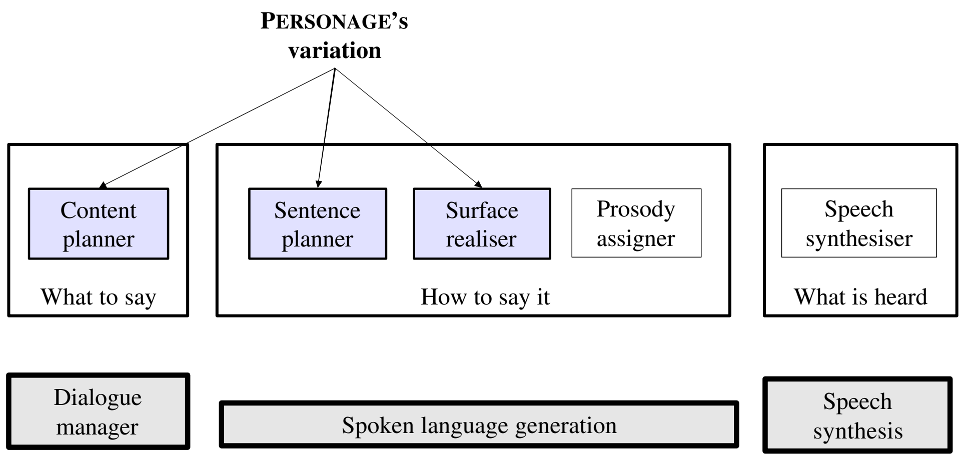


Figure 5.2: The spoken language generation pipelined architecture, with components included in PERSONAGE in the bold inner boxes.

In a typical dialogue system, the output of the realiser is annotated for prosodic information by the prosody assigner, before being sent to the text-to-speech engine to be converted into an acoustic signal. As shown in Figure 5.2, this thesis focuses strictly on linguistic processing and leaves prosody assignment to future work. However, the same methodology could be applied to express personality through prosody as many markers of personality have been identified in speech (see Section 2.2 in Chapter 2). The following sections describe each component in more detail.

5.5 Implementation of generation decisions

While Chapter 4 details the mapping between the psychology findings and potential generation decisions, this section focuses on the full implementation of these parameters in the PERSONAGE generator. Components are described in data processing order according to the architecture in Figure 5.1, together with the mechanisms underlying each generation decision.

5.5.1 Content planning

Relations:	JUSTIFY (N:1, S:2); JUSTIFY (N:1, S:3); JUSTIFY (N:1, S:4); JUSTIFY (N:1, S:5); JUSTIFY (N:1, S:6); JUSTIFY (N:1, S:7)
Content:	1. assert(best (*Chanpen Thai*)) 2. assert(is (*Chanpen Thai*, cuisine (*Thai*))) 3. assert(has (*Chanpen Thai*, food-quality (.8))) 4. assert(has (*Chanpen Thai*, atmosphere (.6))) 5. assert(has (*Chanpen Thai*, service (.8))) 6. assert(is (*Chanpen Thai*, price (*24 dollars*))) 7. assert(is (*Chanpen Thai*, location (*Midtown West*)))

Figure 5.3: An example content plan for a recommendation. N = nucleus, S = satellite.

The first step of the generation process is to convert the restaurant's attribute values into a *content plan*, a high level structure reflecting the overall communicative goal of the utterance. In a dialogue system, the initial content plan would be obtained from the dialogue manager. The content plan combines together propositions expressing information about individual attributes using *rhetorical relations* from Mann and Thompson's Rhetorical Structure Theory (RST; 1988). RST is typically used to study the coherence of texts, by recursively defining how multiple spans of text relate to each other. Whenever a span of text is more essential for conveying the desired information, it is referred to as the *nucleus* of the relation (N), whereas the other text span is defined as the *satellite* (S). Two types of communicative goals are supported in PERSONAGE: *recommendation* and *comparison* of restaurants. Figure 5.3 shows an example content plan for a recommendation. The content plan is automatically converted into an equivalent tree structure in Figure 5.4, referred to as the *content plan tree*. Each recommendation content plan contains a claim (nucleus) about the overall quality of the selected restaurant(s), supported by a set of satellite propositions describing their attributes. The propositions—the leaves in the content plan tree—are assertions labelled *assert-attribute(selection name)* in Figure 5.4. Recommendations are characterised by a JUSTIFY rhetorical relation associating the claim with all the other propositions, which are linked together through an INFER relation. The JUSTIFY relation is a mononuclear rhetorical relation in which the satellite supports the speaker's right to express the nucleus. The INFER relation is a multinuclear relation expressing related information without specific constraints.[5]

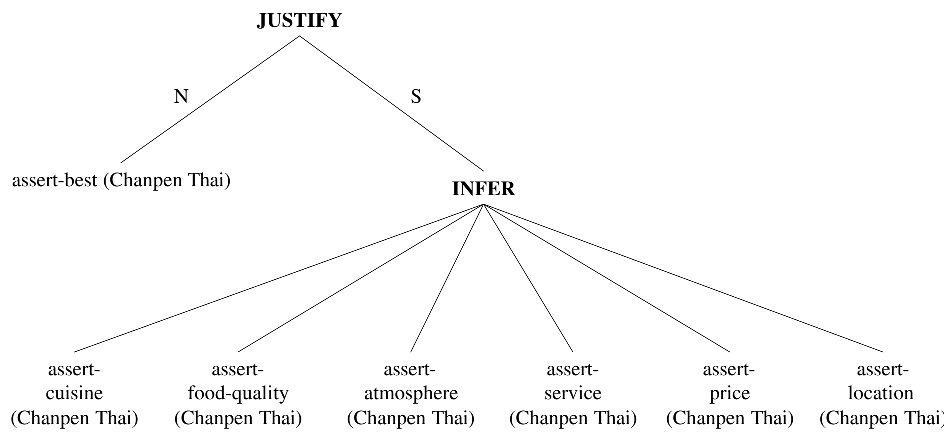


Figure 5.4: An example content plan tree for a recommendation for Chanpen Thai, using all the restaurant attributes. N = nucleus, S = satellite.

In comparisons, the attributes of multiple restaurants are compared using the CONTRAST multinuclear rhetorical relation. This relation combines propositions describing the same attributes for different restaurants, joined together through an INFER relation. An example content plan tree for a comparison between two restaurants is illustrated in Figure 5.5.

The literature presented in Chapter 4 suggests many generation decisions at the content plan level, including parameters influencing the size of the content plan tree, the rhetorical relations used and the polarity of the propositions expressed.

Content size: Extraverts are more talkative than introverts [Furnham, 1990, Pennebaker and King, 1999], although it is not clear whether they actually produce more content, or are just redundant and wordy. Thus various parameters relate to the amount and type of content produced. The VERBOSITY parameter controls the number of propositions selected from the content plan. The parameter value defines the ratio of propositions that are kept in the final content plan tree, while satisfying constraints dependent on the communicative goal: a recommendation must include a claim, and a comparison must include a pair of contrasted propositions. Whereas the VERBOSITY parameter defines the number of propositions in the

[5]The INFER relation is similar to the JOINT relation in the RST literature.

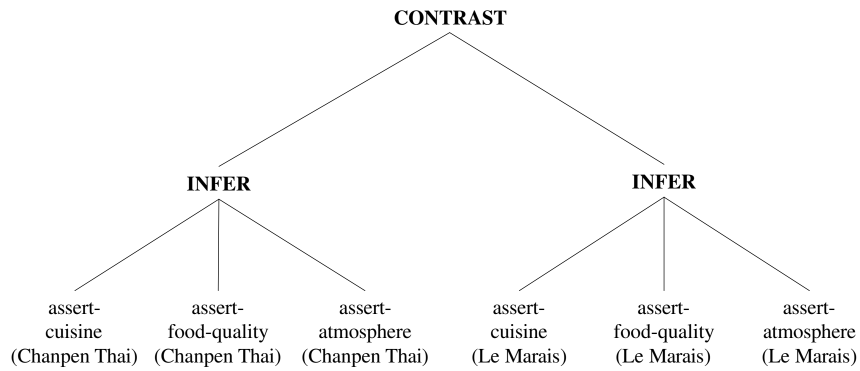


Figure 5.5: An example content plan tree for a comparison between Chanpen Thai and Le Marais, using three attributes. All relations are multinuclear.

final content plan, parameters controlling polarity determine what propositions are selected (see below).

The REPETITION parameter adds an exact repetition: the proposition node is duplicated and linked to the original content by a RESTATE rhetorical relation. The continuous parameter value (between 0 and 1) is mapped linearly to the number of repetitions in the content plan tree, i.e. between 0 and a domain-specific maximum (set to 2 in our domain). Rather than copying the existing proposition, the RESTATEMENT parameter adds a paraphrase to the content plan, obtained from the generation dictionary (see Section 5.5.2). If no paraphrase is found, one is created automatically by substituting content words with the most frequent WordNet synonym (see Section 5.5.5).

Polarity: Extraverts are more positive; introverts are characterised as engaging in more 'problem talk' and expressions of dissatisfaction [Thorne, 1987]. To control for polarity, propositions are defined as positive or negative based on the scalar rating of the corresponding attribute, normalised between 0 and 1. The claim in a recommendation is assigned a maximally positive polarity of 1, whereas the *cuisine* and *location* attributes have a neutral polarity, i.e. a domain-dependent constant set to .58 for our restaurant database.[6] Given a selected restaurant, all other attributes are defined as negative or positive depending on whether their normalised scalar value is below or above the neutral point. There are multiple parameters associated with polarity. The CONTENT POLARITY parameter controls whether the content is

[6]This neutral value was chosen based on the perception of Zagat Survey's restaurant ratings, i.e. scores below $\frac{17.5}{30} = .58$ are considered negative.

mostly negative (e.g. *'Chanpen Thai has mediocre food'*), neutral (e.g. *'Le Marais is a French restaurant'*), or positive (e.g. *'Babbo has fantastic service'*). If there is enough polarised content given the required content plan tree size, the following propositions are selected depending on the input CONTENT POLARITY parameter:

CONTENT POLARITY	$\in [0, .25[$	only negative propositions
	$\in [.25, .5[$	negative and neutral propositions
	$\in [.5, .75[$	neutral and positive propositions
	$\in [.75, 1]$	only positive propositions

If there are not enough propositions in the resulting set to satisfy the verbosity constraint, propositions with the closest polarity are added until the required content plan size is reached. Additionally, a constraint requiring that a comparison content plan tree contains at least one CONTRAST relation is enforced, thus the tree is likely to include propositions with different polarities.

From the filtered set of propositions, the POLARISATION parameter determines whether the final content includes attributes with extreme scalar values or not (e.g. *'Chanpen Thai has fantastic staff'* vs. *'Chanpen Thai has decent staff'*). The final content plan tree therefore contains the propositions whose normalised distance to the neutral point is the closest to the target POLARISATION value ($\in [0, 1]$), while its size is defined by the VERBOSITY constraint.

In addition, polarity can also be implied more subtly through rhetorical structure. The CONCESSIONS parameter controls the way in which negative and positive information is presented, i.e. whether two propositions with different polarity are presented objectively, or if one is foregrounded and the other backgrounded. If two opposed propositions are selected for a concession, a CONCEDE mononuclear rhetorical relation is inserted between them. More precisely, the parameter controls the ratio of concessions being inserted out of all proposition pairs with opposite polarity.[7] While the CONCESSIONS parameter captures the tendency to put information into perspective, the CONCESSION POLARITY parameter controls whether the positive or the negative content is conceded, i.e. marked as the satellite of the CONCEDE relation (e.g. *'even if the food is good, it's expensive'* vs. *'even if the food is expensive, it's good'*). The parameter value determines the ratio of positive concessions out of the total number of concessions in the content plan tree.

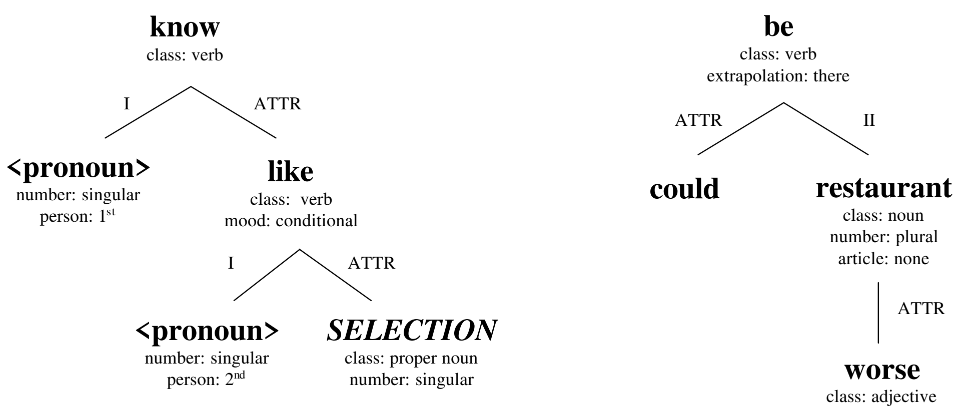
[7]We only consider concessions between attributes of the same restaurant.

Content ordering: Although extraverts use more positive language [Thorne, 1987, Pennebaker and King, 1999], it is unclear how they position the positive content within their utterances. Additionally, the position of the claim affects the persuasiveness of an argument [Carenini and Moore, 2000]. The POSITIVE CONTENT FIRST parameter therefore controls whether positive propositions—including the claim—appear first or last, i.e. the order in which the propositions are aggregated. The parameter controls the ratio of sibling proposition pairs that are ordered with increasing polarity. Although this parameter determines the ordering of the nodes of content plan tree, some aggregation operations can still impose a specific ordering (e.g. BECAUSE CUE WORD to realise the JUSTIFY relation, see Section 5.5.3).

While the INITIAL REJECTION, REQUEST CONFIRMATION and COMPETENCE MITIGATION parameters can also be seen as content planning parameters, they are modelled at the pragmatic marker insertion level as they only affect the beginning of the utterance (see Section 5.5.4).

5.5.2 Syntactic template selection

Once the content planner has determined *what* will be talked about in the utterance, the remaining components control *how* the information is to be conveyed. The first sentence planning component associates each proposition in the content plan with a syntactic template. PERSONAGE manipulates syntactic templates referred to as Deep Syntactic Structures (DSyntS). DSyntS are syntactic representations inspired by Melčuk's Meaning-Text Theory [1988], a linguistic framework in which language is modelled as a multi-stage rule-based process, that gradually modifies the utterance representation from semantics to text. DSyntS can be converted to a text string using the RealPro surface realiser [Lavoie and Rambow, 1997]. The templates are stored in a small handcrafted generation dictionary, containing 18 DSyntS: 12 for the recommendation claim and one per attribute. The DSyntS can contain variables that are filled at generation time, such as the restaurant's name or cuisine. Figure 5.6 shows two DSyntS expressing the recommendation claim. The DSyntS selection process assigns each candidate DSyntS to a point in a three-dimensional space, characterising the DSyntS' syntactic complexity, number of self-references and polarity. Parameter values are normalised over all candidate DSyntS, so the DSyntS closest to the target values can be computed.



(a) *I know you would like SELECTION* (b) *There could be worse restaurants*

Figure 5.6: Two example DSyntS for a recommendation claim. The lexemes are in bold, and the attributes below indicate non-default values in the RealPro realiser. Branch labels indicate dependency relations, i.e. I = subject, II = object and ATTR = modifier. Lexemes in italic are variables that are instantiated at generation time.

Syntactic complexity: Furnham [1990] suggests that introverts produce more complex constructions: the SYNTACTIC COMPLEXITY parameter controls the number of subordinate clauses of the DSyntS chosen to represent the claim, based on Beaman's definition of syntactic complexity [1984].[8] For example, the claim in Figure 5.6(a) is rated as more complex than the one in Figure 5.6(b), because the latter has no subordinate clause.

Self-references: As extraverts and neurotics make more self-references [Pennebaker and King, 1999], the SELF-REFERENCES parameter controls whether the claim is made in the first person, based on the speaker's own experience, or whether the claim is reported as objective or information obtained elsewhere. The SELF-REFERENCES value is computed from the DSyntS by counting the number of first person pronouns. For example, the template in Figure 5.6(a) contains one self-reference, while the template in Figure 5.6(b) does not.

Polarity: While polarity can be expressed by content selection and structure, it can also be directly associated with the DSyntS. The TEMPLATE POLARITY para-

[8] The syntactic complexity is computed as the number of verb nodes in the DSyntS, which is equivalent to the number of subordinate clauses in the final utterance.

meter determines whether the claim has a positive or negative connotation. While automated methods for opinion extraction could be used in the future to annotate the generation dictionary [Pang et al., 2002, Hu and Liu, 2004, Higashinaka et al., 2007], DSyntS are manually annotated for polarity to avoid the introduction of noise due to imperfect opinion modelling. An example claim with low polarity can be found in Figure 5.6(b), i.e. *'There could be worse restaurants'*, whereas the claim in Figure 5.6(a) is rated more positively.

5.5.3 Aggregation

RST relation	Aggregation operations
JUSTIFY	WITH CUE WORD, RELATIVE CLAUSE, SO CUE WORD, BECAUSE CUE WORD, SINCE CUE WORD, PERIOD
CONTRAST	MERGE, HOWEVER CUE WORD, WHILE CUE WORD, CONJUNCTION, BUT CUE WORD, ON THE OTHER HAND CUE WORD, PERIOD
INFER	MERGE, WITH CUE WORD, RELATIVE CLAUSE, ALSO CUE WORD, CONJUNCTION, PERIOD
CONCEDE	EVEN IF CUE WORD, ALTHOUGH CUE WORD, BUT/THOUGH CUE WORD
RESTATE	CONJUNCTION, MERGE WITH COMMA, OBJECT ELLIPSIS

Table 5.1: Clause combining operations for different rhetorical relations, based on SPaRKy's operations [Stent et al., 2004, Walker et al., 2007].

The role of the aggregation component is to combine syntactic templates together into a larger syntactic structure, by associating each pair of sibling propositions in the content plan tree with a *clause-combining operation* that determines how the parent rhetorical relation is to be expressed. For example, poor food quality can be contrasted with good atmosphere using cue words such as 'however' or 'but'. For each rhetorical relation in the content plan tree, the aggregation process randomly selects a clause-combining operation according to the probability distribution for that relation defined by the input aggregation parameters, e.g. the distributions for the INFER relation in Figure 5.3. The aggregation process then selects pairs of operation arguments among the children propositions, until the two associated DSyntS satisfy the constraints of the clause-combining operation, e.g. the MERGE operation requires that both argument DSyntS have the same main verb. If none of the pairs satisfy the constraints, another clause-combining operation is chosen according to the input probability distribution. The aggregation process is guaranteed to terminate as each rhetorical relation implements at least one clause-combining operation with no constraint on the DSyntS, i.e. the PERIOD

Operation	Relations	Description	Sample 1st arg	Sample 2nd arg	Result
MERGE	INFER or CONTRAST	Two clauses can be combined if they have identical verbs and identical arguments and adjuncts except one. The non-identical arguments are coordinated.	Chanpen Thai has good service.	Chanpen Thai has good food quality.	Chanpen Thai has good service and good food quality.
WITH CUE WORD	JUSTIFY or INFER	Two clauses with identical subject arguments can be identified if one of the clauses contains the verb *to have*. The possession clause undergoes *with*-participial clause formation and is attached to the non-reduced clause.	Chanpen Thai is a Thai restaurant.	Chanpen Thai has good food quality.	Chanpen Thai is a Thai restaurant, with good food quality.
RELATIVE CLAUSE	JUSTIFY or INFER	Two clauses with an identical subject can be identified. One clause is attached to the subject of the other clause as a relative clause.	Chanpen Thai has good food quality.	Chanpen Thai is located in Midtown West.	Chanpen Thai, which is located in Midtown West, has good food quality.
CONJUNCTION	JUSTIFY, INFER or CONTRAST	Two clauses are conjoined with a coordinating conjunction. They are separated by a comma if the right clause already contains a conjunction.	Chanpen Thai has good food quality.	Chanpen Thai has good service.	Chanpen Thai has good food quality and it has good service.
ON THE OTHER HAND CUE WORD	CONTRAST	Combines clauses by inserting a cue word at the start of the second clause, resulting in two separate sentences.	Chanpen Thai has very good decor.	Baluchi's has mediocre decor.	Chanpen Thai has very good decor. On the other hand, Baluchi's has mediocre decor.
EVEN IF CUE WORD	CONCEDE	Combines clauses by inserting the *even if* adverbial at the start of the satellite clause. The order of the arguments is determined by the order of the nucleus (N) and the satellite (S), yielding two distinct operations, EVEN IF CUE WORD NS and EVEN IF CUE WORD SN.	Chanpen Thai has very good decor.	Chanpen Thai's has mediocre food quality.	Chanpen Thai has very good decor, even if it has mediocre food quality.
MERGE WITH COMMA	RESTATE	Merges repeated clauses in the same way as the MERGE operation, but ensures that the non-identical arguments are separated by a comma.	Chanpen Thai has very good service.	Chanpen Thai has fantastic waiters.	Chanpen Thai has very good service, fantastic waiters.
OBJECT ELLIPSIS	RESTATE	Coordinates clauses and replaces the object of the first clause by a three-dot ellipsis.	Chanpen Thai has very good service.	Chanpen Thai has fantastic waiters.	Chanpen Thai has... It has fantastic waiters.
PERIOD	Any	Two clauses are joined by a period.	Chanpen Thai is a Thai restaurant, with good food quality.	Chanpen Thai has good service.	Chanpen Thai is a Thai restaurant, with good food quality. It has good service.

Table 5.2: Clause-combining operations and examples as described in previous work on the SPARKY generator [Stent et al., 2004, Walker et al., 2007], together with new operations specific to PERSONAGE.

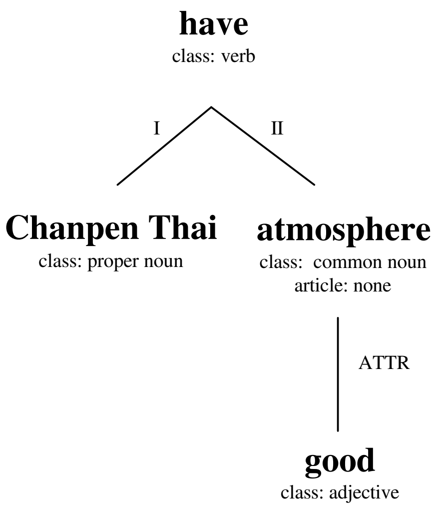
operation, which keeps both argument DSyntS in separate sentences. PERSONAGE uses the SPARKY clause-combining operations [Stent et al., 2004], with additional operations for the RESTATE and CONCEDE rhetorical relations. Table 5.1 shows some of the available operations for each rhetorical relation; their effect on the final utterance is illustrated in Table 5.2.

Aggregation parameters for the INFER relation	**Introvert distribution**	**Extravert distribution**
INFER - MERGE	.20	.50
INFER - RELATIVE CLAUSE	.40	.00
INFER - WITH CUE WORD	.30	.10
INFER - ALSO CUE WORD	.00	.10
INFER - CONJUNCTION	.00	.29
INFER - PERIOD	.10	.01

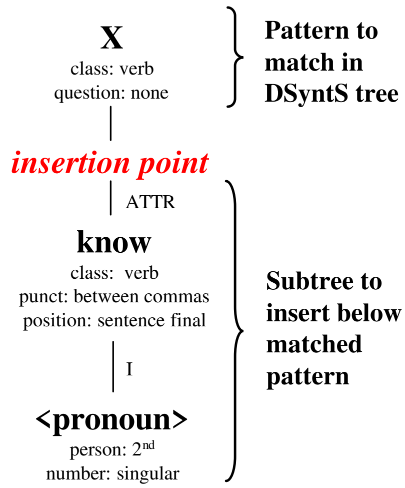
Table 5.3: Probability distribution of aggregation operations expressing the INFER relation for the introvert and extravert parameter settings.[9] Parameter values must add up to 1.

Psychology studies detailed in Chapter 4 show that personality affects the aggregation process. The probability of the operations biases the production of complex clauses, full stops and formal cue words for introverts, to express their preference for complex syntactic constructions, long pauses and rich vocabulary [Furnham, 1990]. Thus, the introvert parameters favour operations such as RELATIVE CLAUSE and PERIOD for the INFER relation, HOWEVER CUE WORD for CONTRAST, and ALTHOUGH CUE WORD for CONCEDE, that we hypothesise to result in more formal language. Extravert aggregation produces longer sentences with simpler constructions and informal cue words. Thus extravert utterances tend to use operations such as a CONJUNCTION to realise the INFER and RESTATE relations, and the EVEN IF CUE WORD for CONCEDE relations. Aggregation parameter values for expressing the INFER relation are illustrated in Table 5.3, for both introvert and extravert parameter settings. Aggregation settings for other traits are based on the findings detailed in Chapter 4.

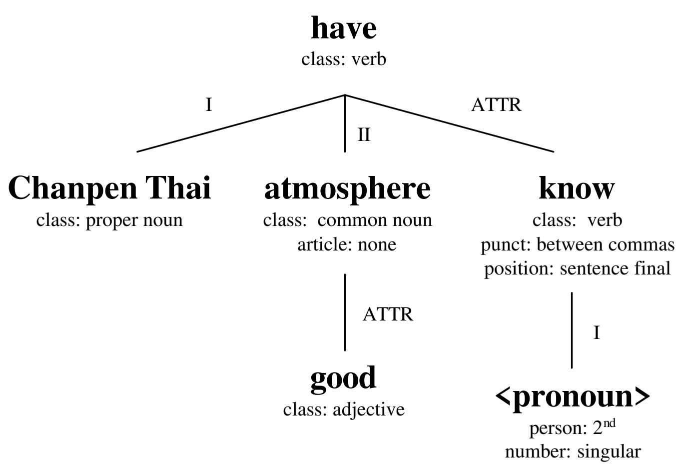
[9] The input selection probability does not entirely reflect the probability that an operation will appear in the output utterance, as the latter is also dependent on the constraints the operation imposes on its DSyntS arguments. For example, the MERGE operation requires both DSyntS to have the same verb, while the CONJUNCTION operation does not. Thus, individual probabilities are scaled to counterbalance these constraints.



(a) Example input DSyntS realised as *'Chanpen Thai has good atmosphere'*.



(b) Syntactic representation of the insertion constraints for the pragmatic marker *you know*.



(c) Modified DSyntS after the insertion of the pragmatic marker below the main verb matching the pattern defined in Figure 5.7(b)'s root node.

Figure 5.7: Illustration of the pragmatic marker insertion process for the hedge *you know* in the DSyntS *'Chanpen Thai has good atmosphere'*.

5.5.4 Pragmatic marker insertion

Many personality markers identified in Chapter 4 are not related to content selection or structuring, rather they manifest themselves through localised syntactic elements reflecting pragmatic effects that only affect a small part of the utterance. To control the insertion of such markers, PERSONAGE implements a pragmatic marker insertion component. A handcrafted database contains syntactic elements characterising each pragmatic marker. For each marker, the insertion process involves traversing the aggregated DSyntS to identify *insertion points* satisfying the syntactic constraints specified in the database. Figure 5.7 illustrates the matching and insertion process for the hedge *you know*. Each entry in the marker database consists of a syntactic pattern to be matched in the DSyntS, such as the root node in Figure 5.7(b), and an insertion point element corresponding to the location in the DSyntS where to insert the subtree representing the marker. Given the input DSyntS in Figure 5.7(a) *'Chanpen Thai has good atmosphere'*, the verb *to have* is matched with the root node of the structure in Figure 5.7(b), and thus the subtree below the insertion point is inserted under Figure 5.7(a)'s root node. The resulting DSyntS is in Figure 5.7(c), realised as *'Chanpen Thai has good atmosphere, you know'*. The utterance is modified at the syntactic level rather than at the surface level, to reduce the complexity of each operation by relying on the surface realiser for grammaticality. For example, pragmatic markers are added without controlling the final word order, while positional constraints can be enforced when required, e.g. the *position* attribute in Figure 5.7(b) specifies that *you know* should be in sentence final position. Similarly, while the *punct* attribute specifies that the marker must appear between commas—irrespectively of its position in the utterance, the realiser ensures that the sentence is punctuated correctly by removing commas preceding the final full stop.

Syntactically embedded markers: PERSONAGE implements a binary generation parameter for most pragmatic markers listed in Tables 5.4 and 5.5, using the insertion mechanism detailed in the previous paragraph. At generation time, syntactic patterns are randomly chosen (with a uniform distribution) among markers with parameter values set to 1, and matched against the aggregated DSyntS. The insertion process ends when there are no markers left in the database, or when the number of successful insertions is above a constant threshold (heuristically set to 5 for the current domain) to avoid producing unnatural utterances.

Other markers: While most pragmatic markers are implemented as described above, additional markers require more complex syntactic processing and are implemented independently.

Referring expression generation is a traditional problem in NLG [Reiter and

Marker	Constraints	Example
General markers:		
NEGATION*	adjective modifier with antonym	Chanpen Thai doesn't have bad atmosphere
EXCLAMATION	sentence-final punctuation	Chanpen Thai has good atmosphere!
IN-GROUP MARKER	clause-final adjunct; available markers are *pal*, *mate* and *buddy*	Chanpen Thai has good atmosphere pal
SUBJECT IMPLICITNESS*	requires a DSyntS of the form *NOUN has ADJ NOUN*	The atmosphere is good
TAG QUESTION*	none	Chanpen Thai has good atmosphere, doesn't it?
STUTTERING*	selection name	Ch-Chanpen Thai has good atmosphere
EXPLETIVES	adjective modifier (*damn*, *bloody*)	Chanpen Thai has damn good atmosphere
	clause-initial adjunct (*oh god*)	Oh god Chanpen Thai has good atmosphere
NEAR EXPLETIVES	adjective modifier (*darn*)	Chanpen Thai has darn good atmosphere
	clause-initial adjunct (*oh gosh*)	Oh gosh Chanpen Thai has good atmosphere
REQUEST CONFIRMATION*	none	You want to know more about Chanpen Thai? Let's see... Chanpen Thai Let's see what we can find on Chanpen Thai Did you say Chanpen Thai?
INITIAL REJECTION*	none	I don't know I'm not sure I might be wrong
COMPETENCE MITIGATION	main verb is subordinated to new clause (*everybody knows that* and *I thought everybody knew that*)	Everybody knows that Chanpen Thai has good atmosphere
	clause-initial adjunct (*come on*)	Come on, Chanpen Thai has good atmosphere
Softeners:		
KIND OF	adjective modifier	Chanpen Thai has kind of good atmosphere
SORT OF	adjective modifier	Chanpen Thai has sort of good atmosphere
SOMEWHAT	adjective modifier with verb *to be*	The atmosphere is somewhat good
QUITE	adjective modifier	Chanpen Thai has quite good atmosphere
RATHER	adjective modifier	Chanpen Thai has rather good atmosphere
AROUND	numeral modifier	Chanpen Thai's price is around $44
SUBORDINATE	main verb is subordinated to new clause; available clauses are *I think that* and *it seems (to me) that*	It seems to me that Chanpen Thai has good atmosphere
Filled pauses:		
LIKE	verb modifier	Chanpen Thai has, like, good atmosphere
ERR	clause-initial adjunct	Err... Chanpen Thai has good atmosphere
MMHM	clause-initial adjunct	Mmhm... Chanpen Thai has good atmosphere
I MEAN	clause-initial adjunct	I mean, Chanpen Thai has good atmosphere
YOU KNOW	clause-final adjunct	Chanpen Thai has good atmosphere, you know

Table 5.4: Pragmatic markers implemented in PERSONAGE, with insertion constraints and example realisations. An asterisk indicates that the pragmatic marker requires specific processing and was not implemented through pattern matching and insertion.

Marker	Constraints	Example
Emphasisers:		
REALLY	adjective modifier	Chanpen Thai has really good atmosphere
BASICALLY	clause-initial adjunct	Basically, Chanpen Thai has good atmosphere
ACTUALLY	clause-initial adjunct	Actually, Chanpen Thai has good atmosphere
JUST	pre-verbal modifier of *to have*	Chanpen Thai just has good atmosphere
	post-verbal modifier of *to be*	The atmosphere is just good
Acknowledgment markers:		
YEAH	clause-initial adjunct	Yeah, Chanpen Thai has good atmosphere
WELL	clause-initial adjunct	Well, Chanpen Thai has good atmosphere
OH	clause-initial adjunct	Oh, Chanpen Thai has good atmosphere
RIGHT	clause-initial adjunct	Right, Chanpen Thai has good atmosphere
OK	clause-initial adjunct	Ok, Chanpen Thai has good atmosphere
I SEE	clause-initial adjunct	I see, Chanpen Thai has good atmosphere

Table 5.5: Pragmatic markers implemented in PERSONAGE (second part), with insertion constraints and example realisations. An asterisk indicates that the pragmatic marker requires specific processing and was not implemented through pattern matching and insertion.

Dale, 2000], which is solved in PERSONAGE by pronominalising any occurrence of a restaurant name following a reference to the same selection, e.g. *'Chanpen Thai is the best, it has great service'*. However, proximal deictic expressions can be seen as a way to express involvement and empathy [Brown and Levinson, 1987], e.g. *'this restaurant has great service'*. Thus, a PRONOMINALISATION parameter controls whether referring expressions are expressed as personal pronouns or proximal demonstrative phrases, by specifying the ratio of pronouns out of all referring expressions in the utterance. Concerning the implementation, the RealPro surface realiser automatically selects the personal pronoun based on the selection's DSyntS node; inserting a demonstrative phrase requires replacing the selection's lexeme with a generic noun (e.g. *restaurant*) and setting the determiner to a demonstrative.

As negations indicate both introversion and a lack of conscientiousness [Pennebaker and King, 1999, Mehl et al., 2006], a NEGATION parameter inserts a negation while preserving the initial communicative goal. If the parameter is enabled, an adjective modifying a verb or its object is randomly selected from the DSyntS, and its antonym is retrieved from WordNet [Fellbaum, 1998]. If the query is successful, the adjective's lexeme is replaced by the antonym and the governing verb is negated,[10] e.g. *'Chanpen Thai has good atmosphere'* becomes *'Chanpen Thai doesn't*

[10]At the DSyntS level the negation is represented as an attribute of the verb element, the actual

have bad atmosphere'. Adjectives in the domain are manually sense-tagged to ensure that they can be substituted by their antonym. Also, a maximum of one negation can be inserted to prevent the utterance from sounding unnatural.

Heylighen and Dewaele [2002] found that extraverts use more implicit language than introverts. A SUBJECT IMPLICITNESS parameter thus determines whether predicates describing restaurant attributes are expressed with the restaurant's name in the subject, or with the attribute itself by making the reference to the restaurant implicit (e.g. *'Chanpen Thai has good atmosphere'* vs. *'the atmosphere is good'*). The syntactic transformation involves shifting the object attribute to the subject, while promoting the adjective below the main verb, and changing the main verb's lexeme to *to be*. Hence, the transformation requires an input DSyntS matching the template *NOUN has ADJECTIVE NOUN*.

As speech disfluencies are associated with anxiety and neuroticism [Scherer, 1981], a STUTTERING parameter modifies the lexeme of a randomly selected proper noun by repeating the first two letters two or three times, e.g. *'Ch-Ch-Chanpen Thai'*. Only selection names are repeated as they are likely to be new to the speaker, the stuttering can therefore be interpreted as non-pathological. Also, allowing disfluencies to affect any word requires determining what words can be altered, which involves deep psycholinguistic modelling that is beyond the scope of this work.

PERSONAGE also implements politeness markers such as rhetorical questions. The TAG QUESTION parameter processes the DSyntS by (1) duplicating a randomly selected verb and its subject; (2) negating the verb; (3) pronominalising the subject; (4) setting the verb to the interrogative form and (5) appending the duplicated subtree as a sentence-final adjunct, e.g. *'Chanpen Thai has great food'* results in the insertion of *'doesn't it?'*. The duplicated verb is generally not realised,[11] i.e. only the negated auxiliary appears in the tag question. Additionally, whenever the subject is a first person pronoun, the verb is set to the conditional form and a second person pronoun is inserted, producing *'I would recommend Chanpen Thai, wouldn't you?'*. If the tag question insertion is unsuccessful, e.g. due to an extrapolated subject *'there is'*, a default tag question is appended, producing either *'you see?'*, *'alright?'* or *'okay?'*.

The remaining parameters are content level parameters that we consider as

inflection is done by RealPro in the realisation phase.

[11]The verb *to be* is an exception.

pragmatic markers, as they only affect the beginning of the utterance. The first two parameters are implemented by inserting a full DSyntS before the utterance, randomly chosen from a predefined list with a uniform probability.[12] First, the INITIAL REJECTION parameter reduces the level of confidence of the speaker over the utterance's informational content, by beginning the utterance with either *'I don't know'*, *'I'm not sure'* or *'I might be wrong'*. Second, the REQUEST CONFIRMATION parameter produces an implicit confirmation, which both redresses the hearer's positive face through grounding and emphasises the system's uncertainty about the user's request, e.g. *'you want to know more about Chanpen Thai?'*. In order to convey disagreeableness, a COMPETENCE MITIGATION parameter also presents the user's request as trivial by embedding it as a subordinate clause, e.g. *'everybody knows that Chanpen Thai has good service'*. See Table 5.4 for additional example confirmation and competence mitigation DSyntS.

Once PERSONAGE has attempted to insert the pragmatic markers specified by the input parameter setting, the next component selects the final lexical items in the DSyntS.

5.5.5 Lexical choice

Brennan [1996] argues that lexical choice is crucial to successful individual adaptation in dialogue systems. Thus, PERSONAGE allows many different lexemes to be expressed for each content word, depending on input parameter values.

The lexical selection component processes the DSyntS by sequentially modifying each content word. For each lexeme in the DSyntS, the corresponding WordNet synonyms are mapped to a multi-dimensional space defined by the lexeme's length, frequency of use and strength, using machine-readable dictionaries. The values along each dimensions are normalised over the set of synonyms, and the synonym that is the closest to the target parameter values (in terms of Euclidean distance) is selected. Although word-sense disambiguation techniques could be used in the future, content words are manually sense-tagged to ensure that the synonyms are interchangeable in the dialogue domain. Figure 5.8 illustrates the lexical choice process using the word length and word frequency dimensions, re-

[12] The constraint on the maximum number of pragmatic markers in the utterance also affects the insertion probability of the DSyntS.

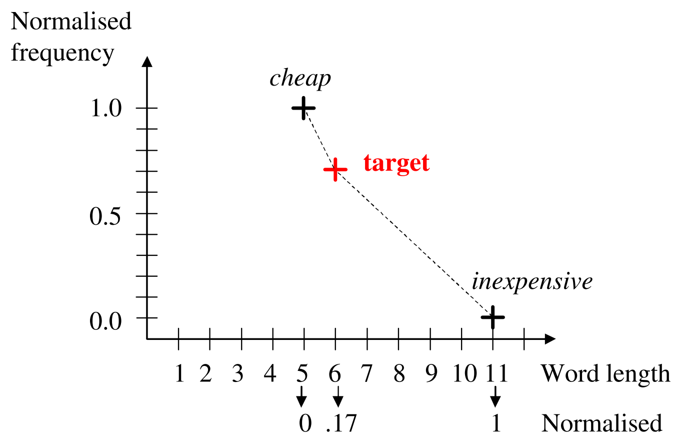


Figure 5.8: Illustration of the lexical selection process between the synonyms *cheap* and *inexpensive* with two input dimensions.

sulting in the selection of *cheap* over *inexpensive* because its length (5 letters) and its normalised frequency (1.0) are closer to the desired target values, i.e. a 6 letter word (normalised length of $\frac{6-5}{11-5} = .17$) with a normalised frequency of .7.

In order to enrich the initial handcrafted pool of synonyms, adjectives extracted by Higashinaka et al. [2007] from a corpus of restaurant reviews and their synonyms are added to the synonym set of each attribute modifier. The list of adjectives is manually filtered for noise. As Higashinaka et al.'s method automatically extracts polarity values for each adjective on a scale from 1 to 5 based on the ratings of the associated reviews, the synonym set for a specific attribute is determined at generation time by mapping the attribute's scalar rating to the polarity scale, e.g. a DSyntS expressing a food quality rating of .42 is mapped to the adjective set with polarity 2 (as $\frac{2}{5} \sim .42$), consisting of the modifiers *bland*, *mediocre* and *bad*. Table 5.6 lists the extracted adjective sets for the food quality attribute, ordered by polarity.

The synonym selection is implemented in PERSONAGE by jointly controlling the average normalised frequency of use, word length and verb strength in each DSyntS.

Frequency of use: Introvert and emotionally stable speakers use a richer vocabulary [Dewaele and Furnham, 1999, Gill and Oberlander, 2003], thus a LEXICON FREQUENCY parameter selects lexical items by their frequency count in the British

Polarity	Adjectives
1	awful, bad, terrible, horrible, horrendous
2	bland, mediocre, bad
3	decent, acceptable, adequate, satisfying
4	good, flavourful, tasty, nice
5	excellent, delicious, great, exquisite, wonderful, legendary, superb, terrific, fantastic, outstanding, incredible, delectable, fabulous, tremendous, awesome, delightful, marvellous

Table 5.6: Adjectives and polarity ratings (5=very positive) for the food quality attribute, extracted from a corpus of restaurant reviews by Higashinaka et al. [2007].

National Corpus.[13]

Word length: As Mehl et al. [2006] show that observers associate long words with agreeableness, conscientiousness and openness to experience, the LEXICON WORD LENGTH parameter controls the number of letters of the selected synonym.

Verb strength: Verb synonyms differ in terms of their connotative strength, such as *appreciate*, *like* and *love*. This variation is controlled in PERSONAGE through the VERB STRENGTH parameter, which orders each verb's synonym set according to the *stronger-than* semantic relation in the VERBOCEAN database [Chklovski and Pantel, 2004]. The process is illustrated in Figure 5.9 for synonyms of the verb *to know*. The ordered synonyms are mapped to equidistant points in the $[0, 1]$ interval to produce the final parameter value, i.e. the weakest verb is associated with 0.0 and the strongest with 1.0. This mapping is based on the assumption that the magnitude of the *stronger-than* relation is constant between contiguous synonyms, i.e. the verb strength is uniformly distributed over the synonym set.

The lexical choice parameters described above associate each candidate synonym with three values, and the one with the closest values to the target is selected. Since values are normalised over the members of the synonym set, all dimensions have the same weight in the selection process.[14] Consider the input DSyntS expressing *'I know you would like Chanpen Thai'*, a low VERB STRENGTH parameter value produces *'I guess you would like Chanpen Thai'*, whereas a high value yields *'I know you would love Chanpen Thai'*. Similarly, a proposition realised as *'this place*

[13] Frequency counts are part-of-speech dependent.

[14] An exception is that verb selection is only affected by the VERB STRENGTH parameter, to ensure that its effect is perceptible in the output utterance.

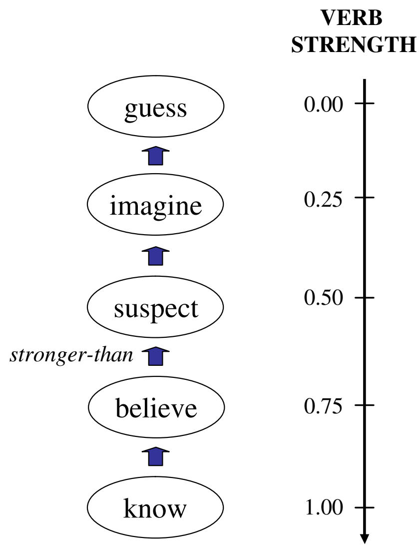


Figure 5.9: Determination of the VERB STRENGTH parameter values for synonyms of the verb *to know*, based on the *stronger-than* semantic relation in VERBOCEAN.

has great ambiance' is converted into *'this restaurant features fantastic atmosphere'* given high LEXICON WORD LENGTH and VERB STRENGTH parameter values together with a low LEXICON FREQUENCY value.

5.5.6 Surface realisation

Surface realisation is the process of converting the DSyntS—a dependency tree—into a sentence string. It can therefore be seen as the reverse of dependency parsing. It involves applying rules of English grammar such as word inflection, function word and punctuation insertion, as well as word ordering. Surface realisation is a fairly well understood process, independent of other components, for which there are re-usable commercially available tools. We use the RealPro surface realiser [Lavoie and Rambow, 1997] to convert the final sequence of DSyntS into a string, with each DSyntS corresponding to one sentence in the utterance.

Even if RealPro's generation decisions are not parameterised in PERSONAGE, some generation decisions presented in Sections 5.5.4 and 5.5.5 can be considered as part of the realisation phase as well, such as the insertion of exclamation marks and lexical choice. As this is only a matter of terminology, surface realisation is defined in this thesis as the application of RealPro's grammatical rules.

5.6 Summary

In this chapter, we have described the overall architecture of the PERSONAGE base generator, as well as the implementation of its generation decisions.

The pipelined NLG architecture presented here has been widely used [Reiter and Dale, 2000], typically in deterministic systems in which linguistic variation is ignored. The contribution of this chapter is to precisely define the implementation of generation decisions derived from psychological studies, to produce a generator that can control this variation. Each parameter implementation represents a hypothesis about how findings in the psychology literature can be modelled computationally, in order to be reproduced in a new, specific dialogue domain.

A secondary goal of this thesis is to project personality independently of the application domain. While we do not evaluate PERSONAGE's re-usability, each generation decision was implemented with the aim to (1) minimise the constraints on the input structure, e.g. by dynamically converting relative input parameter values in the $[0, 1]$ interval into concrete generation decisions, and (2) minimise the level of manual annotation, e.g. by querying machine-readable dictionaries such as WordNet, BNC frequency counts and VERBOCEAN for lexical selection, and by automatically inferring properties of the generation dictionary (i.e. the use of self-references and syntactic complexity).

After mapping findings from the psychology literature to hypothetical generation decisions in Chapter 4, we have shown how to implement them in a base generator, the first building block of our framework. The next chapter evaluates whether human judges recognise the personality of PERSONAGE's utterances generated using the parameter settings suggested in Chapter 4.

Chapter 6

Psychologically Informed Rule-based Generation

The previous chapter details the implementation of generation parameters modelling findings about how people project their personality through language. These personality markers define a variation space of all possible generation decisions. This chapter evaluates how these findings can inform the generator about the region of that space suitable for projecting a target personality. We first evaluate whether the mappings associating personality traits with linguistic markers defined in Chapter 4 can be used to control PERSONAGE's generation process. This approach is referred to as *rule-based*, and it is implemented in the PERSONAGE-RB generator. The following chapters focus on data-driven generation techniques.

6.1 Methodology

In PERSONAGE-RB's rule-based generation approach, each extreme personality type is associated with a set of abstract input parameter values derived from psychological studies, i.e. findings about how a trait is expressed are organised and mapped to generation decisions (see Chapter 4). Parameter values are abstract in the sense that they only indicate trends as opposed to exact generation decisions. Table 6.1 illustrates a subset of this mapping for the extraversion scale, e.g. generating extraversion requires high VERBOSITY, low FILLED PAUSES and low LEXICON WORD LENGTH parameter values, whereas introversion is associated with opposite values (the mappings for all traits are in Tables 4.3 to 4.7 in Chapter 4). The personality model in the rule-based approach can thus produce two extreme personality types for each of the Big Five traits. Table 6.2 shows example output utterances generated from the input content plan in Figure 6.1, for each parameter setting. Additional example outputs can be found in Appendix A.

Relations:	JUSTIFY (nuc:1, sat:2); JUSTIFY (nuc:1, sat:3); JUSTIFY (nuc:1, sat:4); JUSTIFY (nuc:1, sat:5); JUSTIFY (nuc:1, sat:6); JUSTIFY (nuc:1, sat:7)
Content:	1. assert(best (*Chimichurri Grill*)) 2. assert(is (*Chimichurri Grill*, cuisine (*Latin American*))) 3. assert(has (*Chimichurri Grill*, food-quality (.6))) 4. assert(has (*Chimichurri Grill*, service (.6))) 5. assert(has (*Chimichurri Grill*, atmosphere (.4))) 6. assert(is (*Chimichurri Grill*, price (*41 dollars*))) 7. assert(is (*Chimichurri Grill*, location(*Midtown West*)))

Figure 6.1: A content plan for a recommendation.

As it is the base generator's responsibility to ensure that its output is acceptable irrespectively of the input parameter values, parameter trends in Table 6.1 are mapped to extreme parameter values to maximise their impact on the utterance, with *low* = 0.0 and *high* = 1.0 for both continuous and binary parameters. Unspecified parameters are set to default values.[1] An exception is that the probability distributions of aggregation operations are handcrafted for each trait, to factor in the different probabilities of success of clause-combining operations, and to ensure that the parameter values add up to 1 (see Table 5.3 in Chapter 5).

If PERSONAGE-RB were set to express the same trait throughout a dialogue using deterministic rules, every utterance would be generated from the same parameter settings (e.g. neurotic), which could lead to excessive repetitions of identical generation decisions (e.g. hesitancy markers). Parameter values are therefore randomised before generation, according to a normal distribution with a 15% standard deviation around their predefined value,[2] in order to exploit PERSONAGE's variation capabilities.

[1]Default values are chosen to minimise the resulting pragmatic effect, e.g. VERBOSITY and CONTENT POLARITY are set to 0.5, whereas binary pragmatic markers are set to 0.

[2]Binary parameter values are then rounded to 0 or 1.

Introvert findings	Extravert findings	Ref	Parameters	Intro	Extra
Content planning:					
Single topic	Many topics, higher verbal output	1,3,4, 6,13	VERBOSITY	low	high
Strict selection	Think out loud	1*	RESTATEMENTS	low	high
			REPETITIONS	low	high
Problem talk, dissatisfaction, negative emotion words	Pleasure talk, agreement, compliment, positive emotion words	3,14	CONTENT POLARITY	low	high
			REPETITION POLARITY	low	high
			CONCESSION POLARITY	low	high
			POSITIVE CONTENT FIRST	low	high
Not sympathetic	Sympathetic, concerned about hearer (but not empathetic)	10	REQUEST CONFIRMATION	low	high
Syntactic template selection:					
Elaborated constructions	Simple constructions	1*	SYNTACTIC COMPLEXITY	high	low
Problem talk	Pleasure talk	3	TEMPLATE POLARITY	low	high
Aggregation:					
Few conjunctions	Many conjunctions	8	CONJUNCTION, BUT, ALSO CUE WORD	low	high
Many unfilled pauses	Few unfilled pauses	2,7	PERIOD	high	low
Many uses of *although*	Few uses of *although*	9	ALTHOUGH CUE WORD	high	low
Formal language	Informal language	1*,11	RELATIVE CLAUSE	high	low
Pragmatic marker insertion:					
Many nouns, adjectives, prepositions (explicit)	Many verbs, adverbs, pronouns (implicit)	11	SUBJECT IMPLICITNESS	low	high
Many negations	Few negations	3	NEGATION	high	low
Many tentative words (e.g. *maybe*, *guess*)	Few tentative words	3	SOFTENER HEDGES:		
			·SORT OF, SOMEWHAT, QUITE, RATHER, I THINK THAT, IT SEEMS THAT, IT SEEMS TO ME THAT	high	low
Formal language	Informal language	1*,11	·KIND OF, LIKE	low	high
			ACKNOWLEDGMENTS:		
			·YEAH	low	high
			·RIGHT, OK, I SEE, WELL	high	low
Few swear words	Many swear words	6	NEAR EXPLETIVES	low	high
Many unfilled pauses	Few unfilled pauses	2,7	FILLED PAUSES:	high	low
			· ERR, I MEAN, MMHM, YOU KNOW		
Realism	Exaggeration (e.g. *really*)	9*	EMPHASISER HEDGES:		
			·REALLY, BASICALLY, ACTUALLY, JUST	low	high
			EXCLAMATION	low	high
Not sympathetic	Sympathetic, concerned about hearer, minimise positive face threat	10	TAG QUESTION	low	high
Few words related to humans	Many words related to humans (e.g. *man*, *pal*)	12	IN-GROUP MARKER	low	high
Lexical choice:					
Rich vocabulary	Poor vocabulary	1*,4	LEXICON FREQUENCY	low	high
Longer words	Shorter words	6	LEXICON WORD LENGTH	high	low
Realism	Exaggeration	*	VERB STRENGTH	low	high

Table 6.1: Summary of language cues for extraversion (repeated from Chapter 4), as well as the corresponding generation parameters. Asterisks indicate hypotheses, rather than results. Referenced studies are detailed in Table 4.1 in Chapter 4.

Trait	Set	Example output utterance	Score
Extra	low	Chimichurri Grill isn't as bad as the others.	1.00
	high	I am sure you would like Chimichurri Grill, you know. The food is kind of good, the food is tasty, it has nice servers, it's in Midtown West and it's a Latin American place. Its price is around 41 dollars, even if the atmosphere is poor.	6.33
Emot	low	I am not sure! I mean, Ch-Chimichurri Grill is the only place I would recommend. It's a Latin American place. Err... its price is... it's damn ex-expensive, but it pr-pr-provides like, adequate food, though. It offers bad atmosphere, even if it features nice waiters.	4.00
	high	Let's see what we can find on Chimichurri Grill. Basically, it's the best.	6.00
Agree	low	I mean, Chimichurri Grill isn't as bad as the others. Basically, the staff isn't nasty. Actually, its price is 41 dollars. It's damn costly.	2.00
	high	You want to know more about Chimichurri Grill? I guess you would like it buddy because this restaurant, which is in Midtown West, is a Latin American place with rather nice food and quite nice waiters, you know, okay?	5.75
Consc	low	I am not kind of sure pal. Err... Chimichurri Grill is the only place I would advise. It doesn't provide unfriendly service! This restaurant is damn expensive, its price is 41 dollars.	3.00
	high	Let's see what we can find on Chimichurri Grill. I guess you would like it since this eating house, which offers sort of satisfying food and quite satisfactory waiters, is a Latin American eating place.	6.00
Open	low	Err... I am not sure. Mmhm... I mean, Chimichurri Grill offers like, nice food, so I would advise it, also the atmosphere is bad and its price is 41 dollars.	3.50
	high	You want to know more about Chimichurri Grill? I believe you would love it, you know. I guess it's in Midtown West. Although this eating house's price is around 41 dollars, the food is rather satisfactory. This eating place, which provides kind of second-rate atmosphere, is a Latin American restaurant, alright?	5.00

Table 6.2: Example outputs of PERSONAGE-RB for all Big Five dimensions for the input content plan in Figure 6.1, with average judge ratings on the corresponding personality dimension. Personality ratings are on a scale from 1 to 7, with 1 = very low (e.g. introvert) and 7 = very high (e.g. extravert).

6.2 Human evaluation

Our primary hypothesis is that language generated by varying linguistic parameters identified by research on the Big Five can be recognised as expressing the intended trait. To test this hypothesis, expert judges evaluated a set of utterances generated to systematically manipulate parameters related to the extremes of each end of the Big Five traits, e.g. one utterance was generated with the neurotic parameter setting, another with the emotionally stable parameters.

Although there has been considerable work on the expression of various stylistic effects [Hovy, 1988, DiMarco and Hirst, 1993, Paiva and Evans, 2005, Isard et al., 2006, *inter alia*], there have been only a few attempts to evaluate whether the variation produced has the desired pragmatic effect [Fleischman and Hovy, 2002, Porayska-Pomsta and Mellish, 2004]. Most expressions of linguistic variation—e.g. style, emotion, mood and personality—can only be measured subjectively. Thus, a major advantage of the Big Five framework is that it offers standard questionnaires validated over the years by the psychology community [John et al., 1991, Costa and McCrae, 1992, Gosling et al., 2003]. The evaluation of PERSONAGE-RB exploits these questionnaires by using them to have expert judges rate a set of generated utterances as if they had been uttered by a friend responding in a dialogue to a request to recommend restaurants. The judges rate the personality of each utterance by completing the Ten-Item Personality Inventory (TIPI), as this instrument was shown to be psychometrically superior to a 'single item per trait' questionnaire [Gosling et al., 2003]. The items of the questionnaire are listed in Figure 6.2. The answers are averaged to produce a rating for each trait ranging from 1 (e.g. highly neurotic) to 7 (e.g. very stable). To test whether personality can be recognised from a small sample of linguistic output, the judges were asked to evaluate the speaker's personality on the basis of a *single* utterance, i.e. ignoring personality perceptions that could emerge over the course of a dialogue. Additionally, because it was unclear whether the generation parameters defined in Chapter 5 would produce natural sounding utterances, the judges also evaluated the naturalness of each utterance on the same scale.

Because extraversion is the most important of the Big Five traits [Goldberg, 1990], three judges evaluated PERSONAGE-RB in a first experiment focusing strictly on that trait [Mairesse and Walker, 2007]. After positive results were obtained for extraversion, two judges evaluated the four remaining traits in a second experiment. For the sake of clarity, results for both experiments are reported together. The judges consist of researchers and postgraduate students in psychology, history and anthropology. They were all familiar with the Big Five trait theory, but not with natural language generation.

The judges rated a total of 240 utterances, i.e. 80 utterances for the extraver-

Ten-Item Personality Inventory-(TIPI)

Here are a number of personality traits that may or may not apply to you. Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
1	2	3	4	5	6	7

I see myself as:

1. _____ Extraverted, enthusiastic.
2. _____ Critical, quarrelsome.
3. _____ Dependable, self-disciplined.
4. _____ Anxious, easily upset.
5. _____ Open to new experiences, complex.
6. _____ Reserved, quiet.
7. _____ Sympathetic, warm.
8. _____ Disorganized, careless.
9. _____ Calm, emotionally stable.
10. _____ Conventional, uncreative.

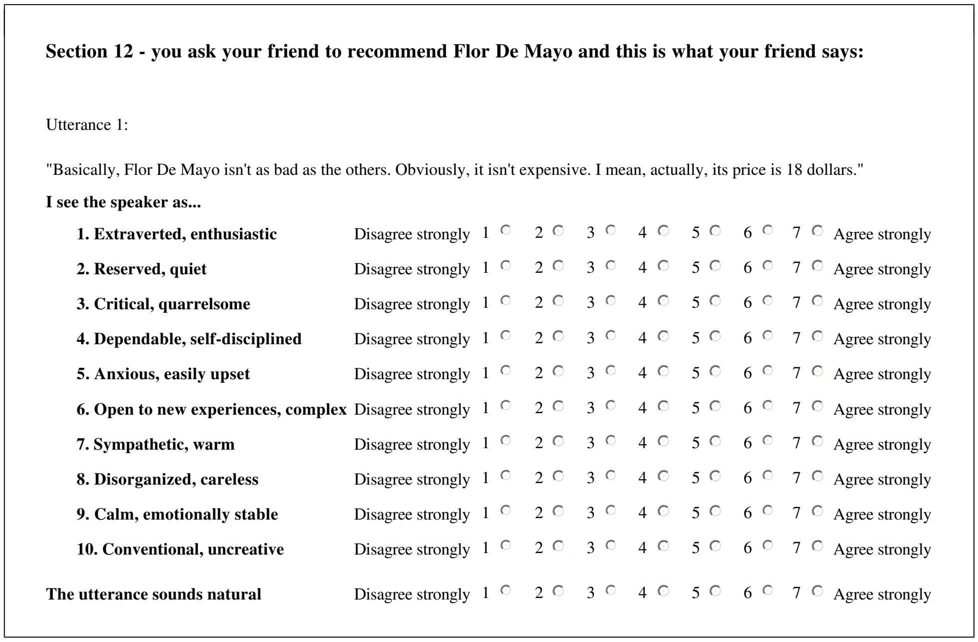
TIPI scale scoring ("R" denotes reverse-scored items):

Extraversion: 1, 6R; Agreeableness: 2R, 7; Conscientiousness; 3, 8R; Emotional Stability: 4R, 9;

Openness to Experiences: 5, 10R.

Figure 6.2: The Ten-Item Personality Inventory (TIPI) scale [from Gosling et al., 2003]. The header was modified in our experiments to evaluate the personality of the speaker, rather than the personality of the judges.

sion experiment and 160 for the evaluation of the other four traits. Utterances were grouped into 20 sets of utterances generated from the same content plan. Each set contained two utterances per trait (four for extraversion), generated with parameter settings for both the low end and the high end of each dimension. These utterances are listed in Appendix A. Table 6.2 shows utterances generated from the content plan in Figure 6.1, with average judge ratings for all dimensions. The judges rated one randomly ordered set at a time, but viewed all utterances in that set before rating them. All questionnaires were filled online; Figure 6.3 shows the online TIPI adapted to the evaluation of personality in our domain. A total of 40



Section 12 - you ask your friend to recommend Flor De Mayo and this is what your friend says:

Utterance 1:

"Basically, Flor De Mayo isn't as bad as the others. Obviously, it isn't expensive. I mean, actually, its price is 18 dollars."

I see the speaker as...

1. Extraverted, enthusiastic	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
2. Reserved, quiet	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
3. Critical, quarrelsome	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
4. Dependable, self-disciplined	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
5. Anxious, easily upset	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
6. Open to new experiences, complex	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
7. Sympathetic, warm	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
8. Disorganized, careless	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
9. Calm, emotionally stable	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
10. Conventional, uncreative	Disagree strongly	1	2	3	4	5	6	7	Agree strongly
The utterance sounds natural	Disagree strongly	1	2	3	4	5	6	7	Agree strongly

Figure 6.3: Online version of the TIPI used in our experiments, adapted to the evaluation of personality in generated utterances.

utterances were rated for each trait (80 for extraversion), with each half targeting one extreme of the dimension. As mentioned in the last section, multiple outputs were generated by allowing each parameter setting to be normally distributed with a 15% standard deviation.

6.3 Results

To assess whether PERSONAGE-RB generates language that can be recognised as conveying extreme personality, Table 6.3 compares the average ratings of the 20 utterances expressing the low end of each trait and the 20 utterances expressing the high end (40 for extraversion). Paired t-tests show that the judges can discriminate between both extreme utterance sets for each trait ($p < .001$). Utterances predicted to be perceived as introvert received an average rating of 2.96 out of 7, whereas utterances predicted to be perceived as extravert received an average rating of 5.98 (difference of 3.02). This difference can also be observed by comparing the distributions of the introvert and extravert utterances in Figure 6.4(a). Emotional stability is the most recognisable trait after extraversion, with a mean rating difference of 2.67 between neurotic and stable utterances. Openness to experience is the hardest trait to convey in our domain, with a rating difference of 1.32 between the utterance sets. This difference, however, is still largely significant ($p < .001$) despite the small number of ratings.

Personality trait	Low	High
Extraversion	2.96	5.98
Emotional stability	3.29	5.96
Agreeableness	3.41	5.66
Conscientiousness	3.71	5.53
Openness to experience	2.89	4.21

Table 6.3: Average personality ratings for the utterances generated with the low and high parameter settings for each trait on a scale from 1 to 7. The ratings of the two extreme utterance sets differ significantly for all traits ($p < .001$, two-tailed).

Inter-rater agreement

Table 6.4 shows that the judges agree significantly for all Big Five traits, although they agree more for some traits than others. The highest agreement is observed for extraversion and emotional stability ($r = .73$ and $r = .67$), and the lowest for conscientiousness and openness to experience ($r = .42$ and $r = .44$). Unsurprisingly, traits that are recognised more accurately produce a higher agreement, suggesting that it is easier to agree on utterances expressing an extreme personality. This level of agreement is only slightly lower than the one observed for conversation extracts in the personality recognition task studied in Chapter 3 ($r = .84$), which is encouraging considering that the judgements presented here are based on a single utterance rather than audio conversation extracts collected over 48 hours.

Personality trait	r
Extraversion	.73
Emotional stability	.67
Agreeableness	.54
Conscientiousness	.42
Openness to experience	.44

Table 6.4: Average inter-rater correlation r over ratings of the utterances generated with the low and high parameter settings for each trait. All correlations are significant at the $p < .05$ level (two-tailed).

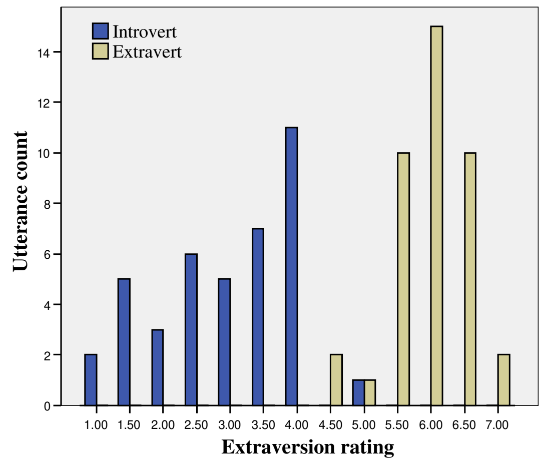
Generation accuracy

While predefined parameters can generate recognisable personality *on average*, the distributions of ratings over the two utterance sets shown in Figures 6.4 and 6.5 give additional insight into the variation produced by PERSONAGE-RB. Reported ratings are averaged over all judges, thus they are less extreme than individual judgements, e.g. an extraversion rating of 1.0 implies that all three judges agreed on that score. As both extremes of each personality dimension are generated, *generation accuracy* is evaluated by splitting ratings into two bins around the neutral rating (4 out of 7), and computing the percentage of utterances with an average rating falling in the bin predicted by its generation parameters. As the rule-based approach presented here aims at producing extreme personality, neutral ratings are considered as misrecognitions.

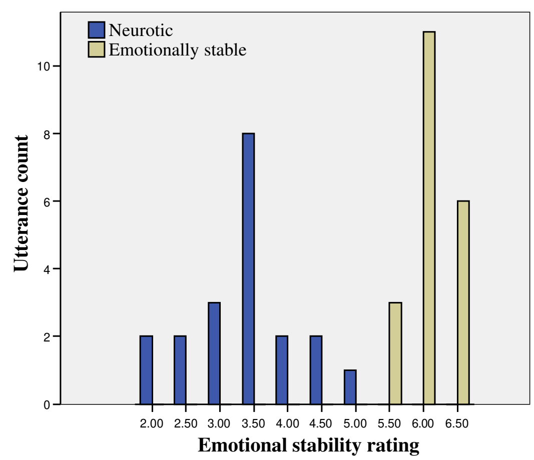
Personality trait	**Low**	**High**	**Overall**
Extraversion	82.5	100.0	91.3
Emotional stability	80.0	100.0	90.0
Agreeableness	70.0	100.0	85.0
Conscientiousness	60.0	100.0	80.0
Openness to experience	90.0	55.0	72.5
All utterances		85.0	

Table 6.5: Generation accuracy (in %) for the utterance sets generated with the low and high parameter settings for each trait. An utterance is correctly recognised if its average rating falls in the half of the scale predicted by its parameter setting. Neutral ratings (4 out of 7) are counted as misrecognitions.

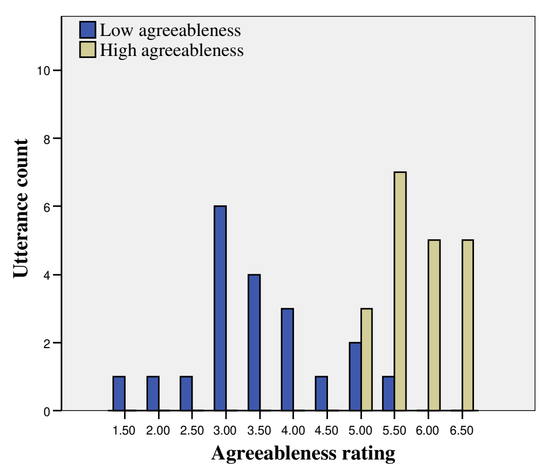
Figure 6.4(a) shows that extravert utterances were all recognised as such, with approximately normally distributed ratings, whereas 17.5% of the introvert utterances were rated as neutral or extravert. Extraversion is the easiest trait to project in our domain, with ratings covering the full range of the scale and an overall accuracy of 91.3% over both utterance sets. Figure 6.4(b) shows that PERSONAGE-RB did not generate utterances perceived as extremely neurotic by all judges, as no utterance were rated below 2 out of 7 on that scale. Also, while all emotionally stable utterances were perceived correctly, 20% of the neurotic utterances were rated as neutral or moderately stable: the ratings' distribution of neurotic utterances is slightly biased towards the positive end of the scale. The parameter settings for agreeableness produce utterances covering the largest range of ratings after extraversion (from 1.5 to 6.5), although 30% of the disagreeable utterances were rated



(a) Extraversion

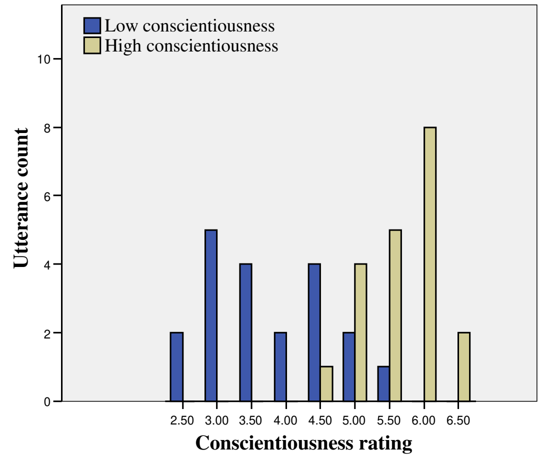


(b) Emotional stability

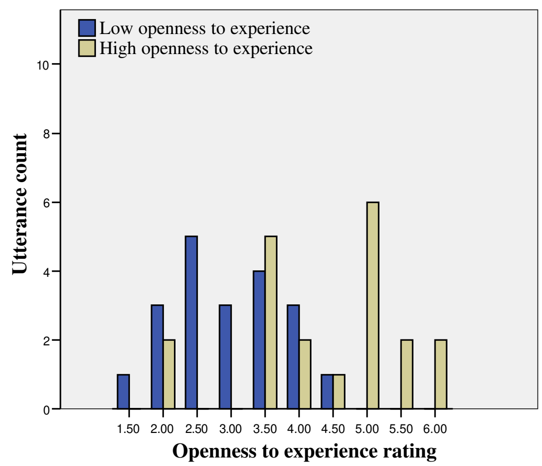


(c) Agreeableness

Figure 6.4: Rating distributions over both extreme utterance sets for extraversion (80 utterances), emotional stability and agreeableness (40 utterances). Ratings are averaged over all judges and rounded to the nearest half-integer.



(a) Conscientiousness



(b) Openness to experience

Figure 6.5: Rating distributions over both extreme utterance sets for conscientiousness and openness to experience (40 utterances each). Ratings are averaged over all judges and rounded to the nearest half-integer.

as neutral or agreeable. See Figure 6.4(c). On the other hand, all agreeable utterances were perceived correctly. Figure 6.5(a) shows that unconscientiousness is more difficult to model, as only 60% of the utterances generated with the corresponding parameter setting were perceived as unconscientious, with no average rating below 2.5 out of 7. However, all conscientious utterances have ratings in the positive end of the scale. Openness to experience is the most difficult dimension to evaluate, as misinterpretations occurred for both ends of the scale (10% for non-open utterances, and 45% for open utterances), yielding an average accuracy of 72.5% for this dimension. Table 6.5 summarises generation accuracies for all traits, showing that PERSONAGE-RB produces an average accuracy of 85%, i.e. a large majority of the utterances were recognised correctly.

Table 6.5 shows that the positive ends of most dimensions are modelled with high precision, while parameter settings for the low ends—typically associated with a low desirability—produce more misrecognised utterances. Openness to experience is the only exception, with a higher accuracy for narrow-minded utterances. This overall trend can be explained by a bias of the judges towards the positive end, as suggested by the overall distributions of ratings. It could also be a consequence of a bias in PERSONAGE-RB's predefined parameter settings, that could be attenuated by recalibrating the parameter values. Finally, it is also possible that some parts of the spectrum of personality cannot be conveyed through language only, or that more than a single utterance is required. After all, the Big Five framework is meant to explain the variance of all aspects of human behaviour, including language as well as gesture, facial expression, social interaction and even long-term personal decisions.

Naturalness

Judges were also asked to evaluate the naturalness of each utterance, i.e. to what extent they could have been uttered by a human. Results in Table 6.6 show that the utterances were seen as moderately natural on average, with a mean rating of 4.59 out of 7. Although naturalness could be improved by adding generation constraints to avoid ungrammatical or inconsistent outputs, we believe it is a promising score given the extreme nature of the modelled personality: utterances produced by a human with such extreme traits may be perceived as unnatural as well. Table 6.7 shows examples of utterances perceived as unnatural.

Personality trait	**Low**	**High**
Extraversion	4.93	5.78
Emotional stability	3.43	4.63
Agreeableness	3.63	5.56
Conscientiousness	3.33	5.33
Openness to experience	3.98	3.85
All utterances	4.59	

Table 6.6: Average naturalness ratings for the utterance sets generated with the low and high parameter settings for each trait.

Utterance	Trait	Nat
You want to know more about Le Rivage and Pintaile's Pizza? I see, right, I would suggest them, you would probably like them, you see? It seems to me that Pintaile's Pizza is inexpensive. Le Rivage is located in Manhattan and it offers quite bad atmosphere, also it's somewhat expensive.	high agreeableness	2.0
You want to know more about Vinnie's Pizza? I see, it's sort of the best restaurant. It's located in Manhattan, isn't it? Although this eating place features unmannerly waiters, the food is somewhat kind of satisfying. Its price is 13 dollars. I suppose this eating house provides terrible ambience.	high openness to experience	1.5
I might be bloody wrong. Err... Edgar's Cafe is the only place that is any good. I mean, this restaurant offers like, bad service, bad waiters, even if the food isn't kind of nasty.	low conscientiousness	1.5

Table 6.7: Example unnatural utterances generated by PERSONAGE-RB. *Nat* = average naturalness rating.

Table 6.6 shows that some traits are perceived as more natural than others. Extravert, agreeable and conscientious utterances are rated as the most natural, with an average rating above 5.5 out of 7. Introvert utterances are also perceived as natural, with ratings close to 5. On the other hand, utterances expressing unconscientiousness, neuroticism, disagreeableness and openness to experience are rated as moderately unnatural, with average scores below 4. A comparison between Tables 6.5 and 6.6 suggests a correlation between naturalness and generation accuracy, however it is not clear whether (1) poor personality recognition is a consequence of unnatural utterances, or whether (2) the projection of inconsistent personality cues causes the low naturalness scores. Although some extreme traits are likely to be perceived as unnatural because they are not commonly observed, this relation suggests that it is important to maintain a high generation quality, by ensuring the plausibility and consistency of the personality markers expressed.

6.4 Summary

The primary contribution of this chapter is to show that personality can be modelled computationally in a language generator, by using findings from psychological studies to control the generation process. We evaluate the mappings associating the findings to concrete generation parameters presented in Chapter 4, providing a systematic framework for testing whether the findings generalise to the application domain.

A secondary contribution is a new, domain-independent method for evaluating personality in natural language generation (NLG), where human judges complete a standard personality questionnaire assessing the personality of the generated utterances as though they were uttered by a human speaker. Our evaluation method shows that the PERSONAGE-RB generator produces recognisable personality (1) in a constrained information presentation domain; (2) through a single utterance; and (3) for all personality traits in the Big Five framework. While previous work has evaluated some aspects of the personality of a small number of handcrafted surface templates [Isbister and Nass, 2000], we do not know of any evaluation of the personality projected by varying parameters at all levels of the NLG pipeline.

Throughout this chapter, personality was considered as a discrete phenomenon, as the parameter settings generate language expressing either the low end or the high end of each personality trait, and only one trait at a time. This capability can be used for dialogue system adaptation in systems supporting a limited range of user models, or other applications that do not require fine-grained variation of the generation output, e.g. artificial characters with static behaviour. However, the wide range of individual differences reflected by the literature on the Big Five [Allport and Odbert, 1936, Norman, 1963, Goldberg, 1990] as well as recent work in medical research [Marcus et al., 2006] suggest that personality varies continuously. This continuity is also reflected by the continuous scales used in personality psychology instruments [John et al., 1991, Costa and McCrae, 1992, Gosling et al., 2003]. Thus, an open issue is whether personality can be generated continuously by producing language targeting any arbitrary value on the Big Five dimensions, e.g. generating an utterance perceived as 60% extravert or 75% neurotic. This issue is addressed in the following chapters.

Chapter 7

Stochastic Generation Capabilities

The previous chapter presents a rule-based generation approach for the generation of extreme personality traits, in which personality is modelled as a discrete phenomenon. We showed that PERSONAGE-RB can generate recognisable personality, for each of the Big Five traits. However, psychologists generally consider personality as continuous [Norman, 1963, Goldberg, 1990, Marcus et al., 2006], therefore it seems likely that personality is best modelled as a complex of continuous variables. Additionally, applications requiring finer-grained linguistic variation might be limited by the finite set of personality types—e.g. two per trait—that can be projected using the rule-based approach.

The rule-based approach, while it specifies the appropriate region of the variation space for generating a specific trait, requires the manual determination of every parameter value. Thus, extending this approach to continuous variation is not tractable. On the other hand, *data-driven* machine learning techniques offer a scalable approach for continuous personality variation, by automatically learning a model of the relation between personality scores and properties of the utterances.

This chapter lays the ground work for establishing whether the PERSONAGE generator can be used to generate personality close to any arbitrary value on the Big Five scales using stochastic methods. Chapters 8 and 9 will then use this chapter as a basis for developing and evaluating two distinct methods for data-driven generation of personality.

Statistical learning algorithms typically learn from examples, i.e. they require a set of pairs mapping example linguistic cues to personality ratings. What kind of linguistic behaviour should be presented to the learning algorithm? A reasonable answer would be to use personality ratings of PERSONAGE's output utterances, since this is the domain in which we want to convey personality.

While we could use the utterances generated from predefined parameter settings in the last chapter, the high correlation between the predefined generation decisions within each utterance set prevents learning algorithms from filtering out irrelevant decisions that are correlated with relevant ones. In other words, because the same linguistic cues are consistently used to convey a given personality, it is not possible to identify which cue—or utterance feature—is responsible for observed discrepancies between the target personality and the judges' ratings. Hence, this chapter studies a new set of *random* utterances generated with uniformly distributed parameter values.

Furthermore, although representing the utterance using appropriate features is an essential aspect of accurate statistical modelling, no feature representation will be effective if the variation space—the set of all utterances that can be generated—does not cover the range of the target variation dimensions, or if it contains mostly unnatural utterances. Therefore, first, Section 7.1 establishes that (1) the distribution of the random utterances covers the whole target variation range, while the distribution of the rule-based utterances only covers the extremes; (2) that the level of agreement between judges of the personality of the randomly generated utterances is high enough to provide a stable population of ratings to use as the target score to be learnt by the learning algorithms employed in Chapters 8 and 9; (3) that the level of naturalness of the randomly generated utterances is high enough to generate stable judgements of personality traits.

As input utterances need to be described in terms of relevant features, Section 7.2 investigates what utterance features provide important information regarding the personality conveyed by the utterance, purely based on the judges' utterance ratings. Thus, this analysis is also an evaluation of the hypotheses made in Chapter 4 regarding the capability of linguistic markers to convey personality in PERSONAGE's information presentation domain.

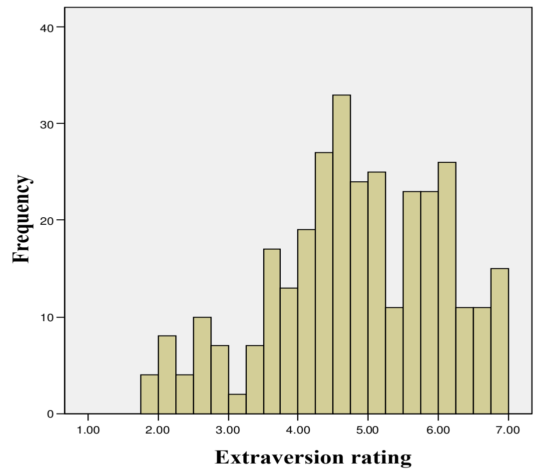
7.1 Generation coverage and quality

The random utterances are rated using the same process as the one detailed in Section 6.2 in Chapter 6, i.e. the same two judges filled the Ten-Item Personality Inventory [Gosling et al., 2003] for a total of 160 random utterances (three judges rated 320 utterances for extraversion), generated from 20 content plans (40 for extraversion). A subset of these utterances can be found in Appendix B. Judges were also asked to evaluate the naturalness of each utterance, i.e. whether or not it could have been uttered by a human.

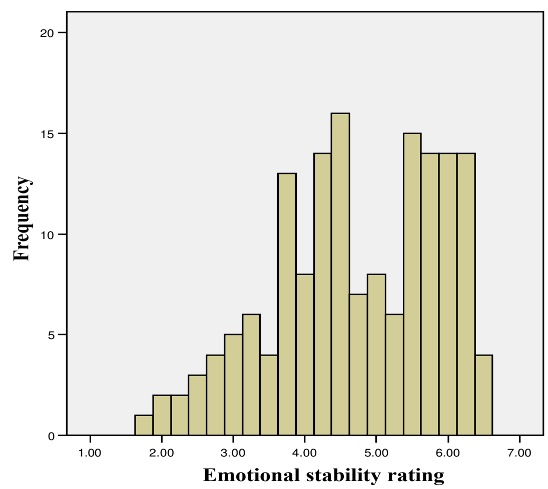
7.1.1 Ratings distribution

While the predefined parameter settings presented in Chapter 4 provide one way to express each personality trait, the randomly generated utterances offer the opportunity to assess whether other parameter settings can be used to project personality. Figures 7.1 and 7.2 illustrate the distribution of the average judge ratings over the 160 random utterances (320 for extraversion), for all Big Five traits. Figure 7.1(a) shows that PERSONAGE can produce utterances covering most of the extraversion scale, with ratings approximately normally distributed over the $[1.75, 7.0]$ interval. No utterance was rated as extremely introvert by all judges (i.e. with a 1.0 score), however four utterances were rated as low as 1.75 out of 7. Generally, a bias towards the positive end is observed, as 73.4% of the utterances are rated as extravert, i.e. with a rating above 4. Concerning emotional stability, Figure 7.1(b) illustrates that PERSONAGE produces utterance rated from 1.75 up to 6.5 on that scale, with 70% of the utterances in the positive half. The coverage of the agreeableness dimension is slightly narrower, with ratings in the $[2.5, 6.5]$ interval,[1] and with 75% of the random utterances perceived as agreeable. See Figure 7.1(c). A similar bias is observed for conscientiousness in Figure 7.2(a), with 67.5% of the utterances perceived as conscientious, and no utterance rated below 2.0 on average. Finally, Figure 7.2(b) shows that PERSONAGE generates utterances symmetrically distributed over the openness to experience scale, with ratings ranging from 1.75 to 6.25.

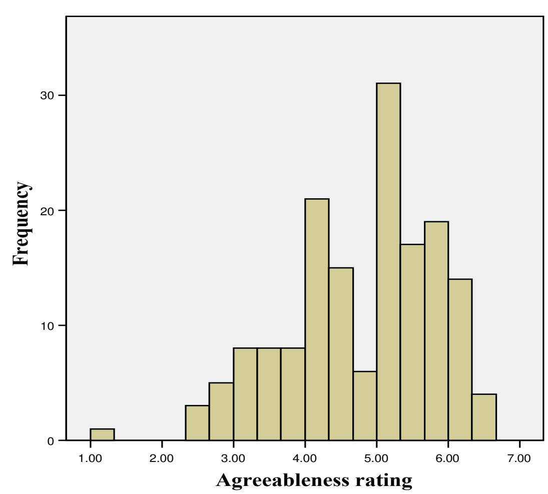
[1] Ignoring one utterance perceived as extremely disagreeable (1.25 score).



(a) Extraversion

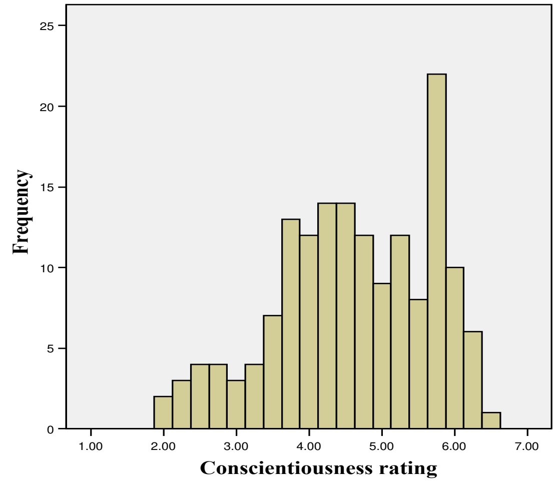


(b) Emotional stability

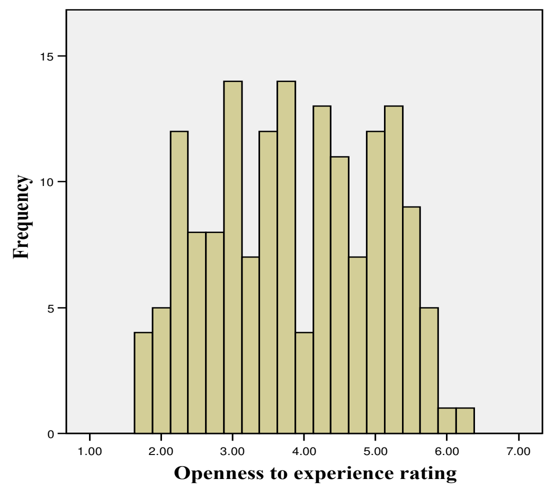


(c) Agreeableness

Figure 7.1: Rating distributions over the random utterance set for extraversion (320 utterances), emotional stability and agreeableness (160 utterances). Ratings are averaged over all judges.



(a) Conscientiousness



(b) Openness to experience

Figure 7.2: Rating distributions over the random utterance set for conscientiousness and openness to experience (160 utterances). Ratings are averaged over all judges.

7.1.1.1 Comparison with the rule-based approach

In order for a data-driven approach to be preferable to the rule-based technique presented in Chapter 6, one must ensure that some utterances are perceived as extreme as those generated from the predefined parameter settings derived from psychological studies. We thus assess PERSONAGE's stochastic generation capabilities by evaluating the *variation range* of the random utterances' personality ratings. When comparing with the rule-based approach—which only generates extreme personality—we evaluate whether randomly generated utterances are perceived as more extreme.

Method	**Rule-based**		**Random**		
Dimension	**Low**	**High**	**Min**	**Avg**	**Max**
Extraversion	2.96	5.98	1.75	4.75	7.00
Emotional stability	3.29	5.96	1.75	4.72	6.50
Agreeableness	3.41	5.66	1.25	4.76	6.50
Conscientiousness	3.71	5.53	2.00	4.61	6.50
Openness to experience	2.89	4.21	1.75	3.86	6.25
Naturalness	4.59		1.00	4.38	7.00

Table 7.1: Comparison between the personality and naturalness ratings of the random utterances (*Min*, *Avg* and *Max*) with the average ratings obtained using the predefined parameter settings evaluated in Chapter 6 (*Low* and *High*). Ratings are averaged over all judges.

Method	**Rule-based**		**Random**	
Trait	**Low**	**High**	**Lowest**	**Highest**
Extraversion	2.96	5.98	3.60 ○	6.23 •
Emotional stability	3.29	5.96	3.05	6.25 •
Agreeableness	3.41	5.66	3.26	6.01 •
Conscientiousness	3.71	5.53	3.11 •	5.93 •
Openness to experience	2.89	4.21	2.28 •	5.48 •

•,○ significant increase (•) or decrease (○) of the variation range over the average rule-based ratings ($p < .05$, two-tailed)

Table 7.2: Pair-wise comparison between the most extreme ratings of the random utterances (*Random*) and the ratings obtained with Chapter 6's predefined parameter settings (*Rule-based*), averaged over 20 content plans. Ratings are averaged over all judges.

Table 7.1 shows that some random utterances are perceived as more extreme than the average rule-based utterance, for both ends of each Big Five trait. Although these results suggest that the random parameters produce enough variation when taking all the utterances in our dataset into account, they do not show whether the random parameters are likely to produce extreme personality *for any content plan*, i.e. the large range of ratings observed in Figures 7.1 and 7.2 might result from a few content plans. Thus, Table 7.2 compares the most extreme ratings obtained among the 8 random utterances generated for each content plan with the ratings of the corresponding rule-based utterance.[2] Paired t-tests over 20

[2]There is one rule-based utterance per content plan for each end of each trait, with twice as many for extraversion. Also, extraversion ratings were collected over two separate experiments, resulting

content plans show that on average the most extreme random utterance is significantly more extreme for the positive end of the extraversion, emotional stability and agreeableness scales, and both ends of the conscientiousness and openness to experience scales ($p < .05$, two-tailed). However, random utterances are not perceived as introvert as those generated using the introvert parameter settings. These results suggest that for any content plan, one can expect as much variation as with handcrafted parameter settings with less than 10 utterances, by randomising PERSONAGE's parameters.

While previous results evaluate PERSONAGE's coverage with a fixed number of utterances, Figures 7.3 and 7.4 compare the variation range obtained with different utterance set sizes. The minimum and maximum ratings are averaged over 100 random subsets of the full set of random utterances, by merging together all content plans. It is important to note that this evaluation approximates the variation within a single content plan with the variation across all content plans in our dataset. Nevertheless, a comparison with the average ratings obtained with the rule-based approach shows that more extreme ratings are generated within 10 utterances, for all traits. A consequence is that given a perfect personality recognition model, such a few number of utterances would be enough to generate utterances as extreme as with the rule-based approach. Additionally, Figures 7.3 and 7.4 suggest that 80 utterances are enough to generate a large variation for all traits, as additional utterances only increase the range marginally.[3]

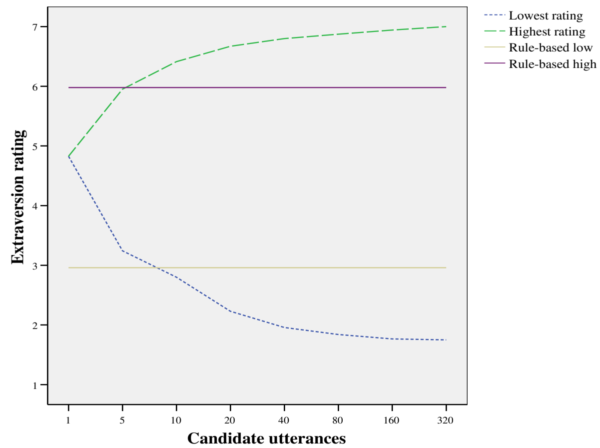
The results presented here suggest that PERSONAGE's coverage is large enough to project fine-grained variation. However, the successful generation of recognisable variation also depends on the overall quality of the utterances. Thus, the next sections evaluate whether the judges agree over the personality of PERSONAGE's random output, as well as whether the utterances are perceived as natural.

7.1.2 Inter-rater agreement

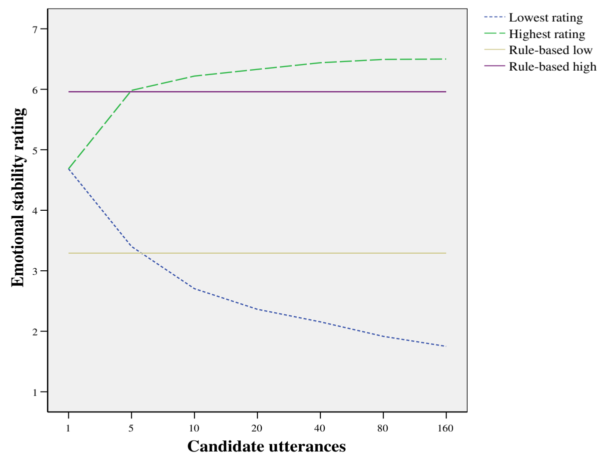
An estimate of the quality of PERSONAGE's output is the *inter-rater agreement*, which reflects whether or not PERSONAGE's personality markers are clear enough to be de-

in a third of the random utterances for that trait being partitioned into 6 utterances per plan, and two thirds into 11 utterances per plan.

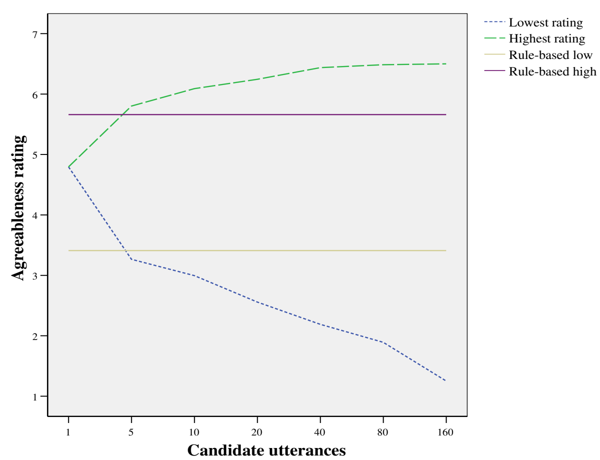
[3]With the exception of agreeableness in Figure 7.3(c), for which the additional utterances produce a substantial range increase.



(a) Extraversion

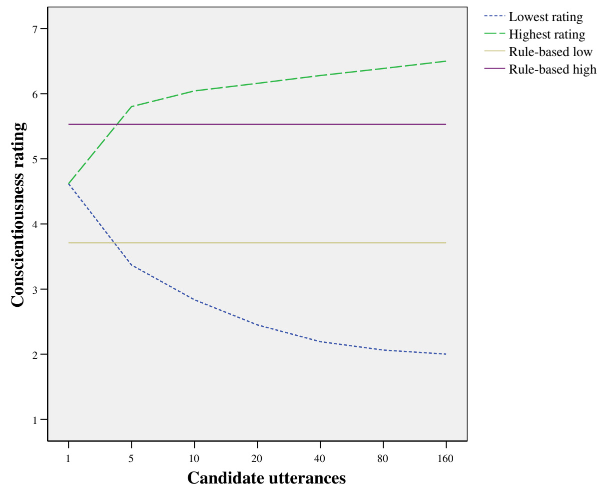


(b) Emotional stability

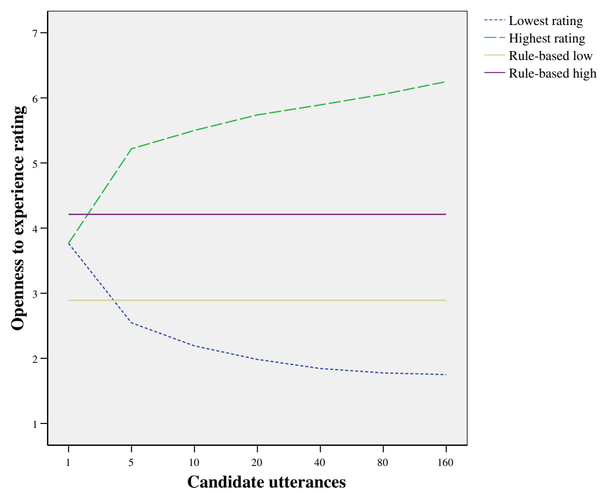


(c) Agreeableness

Figure 7.3: Range of personality scores obtained with different numbers of random utterances, compared with the average rule-based rating with predefined parameter settings. The minimum and maximum ratings are averaged over 100 random utterance subsets. Ratings are averaged over all judges.



(a) Conscientiousness



(b) Openness to experience

Figure 7.4: Range of personality scores obtained with different numbers of random utterances, compared with the average rule-based rating with predefined parameter settings. The minimum and maximum ratings are averaged over 100 random utterance subsets. Ratings are averaged over all judges.

tected by all judges. Table 7.3 shows the average correlations between the judges' ratings over 160 random utterances (320 for extraversion), compared with the 40 utterances generated with predefined parameter settings in Chapter 6 (80 for extraversion).

The judges agree significantly over the personality of the random utterances for all Big Five traits ($p < .05$, two-tailed), with correlations ranging from .26 (conscientiousness) to .40 (agreeableness). However, Table 7.3 shows that the

Parameter set	Rule-based	Random	All
Extraversion	.73	.30	.48
Emotional stability	.67	.33	.39
Agreeableness	.54	.40	.44
Conscientiousness	.42	.26	.32
Openness to experience	.44	.28	.35

Table 7.3: Average inter-rater correlation for the rule-based and random utterances. Correlations under the *All* column were computed over the full dataset.[4] All correlations are significant at the $p < .05$ level (two-tailed).

agreement is lower than on the rule-based utterances. A possible explanation is that the random generation decisions are more likely to produce utterances projecting inconsistent personality cues, which can be interpreted in different ways by the judges. An example of inconsistency can be found in the utterance *'Err... I am sure you would like Chanpen Thai!'*, as it expresses markers of both introversion (filled pause) and extraversion (exclamation mark). A second cause for this difference is that utterances conveying extreme personality are more likely to be agreed on, i.e. the ratings' distributions in Figures 7.1 and 7.2 show that only a few random utterances project extreme personality compared to utterances generated from predefined parameter settings.

However, a low inter-rater agreement does not necessarily imply poor modelling performance: it is a consequence of the difficulty of the task, as well as a result of individual differences in personality recognition. As statistical models are trained on the *average* ratings, the learning algorithm is likely to learn associations that generalise across judges.

7.1.3 Naturalness

As the main objective of a language generator is to produce high quality outputs, we evaluate the effect of stochastic generation on output quality, as perceived by the judges.

Figure 7.5 shows that naturalness ratings are approximately normally distributed over the 320 random utterances, with 67.8% of the utterances rated as natural (rating above or equal to 4) and an average rating of 4.38 out of 7. The bottom row of Table 7.1 shows that the random utterances are rated as slightly less natural

[4]Correlations over the full dataset (All) include cross-trait judgements such as extraversion ratings for utterances with neurotic parameters.

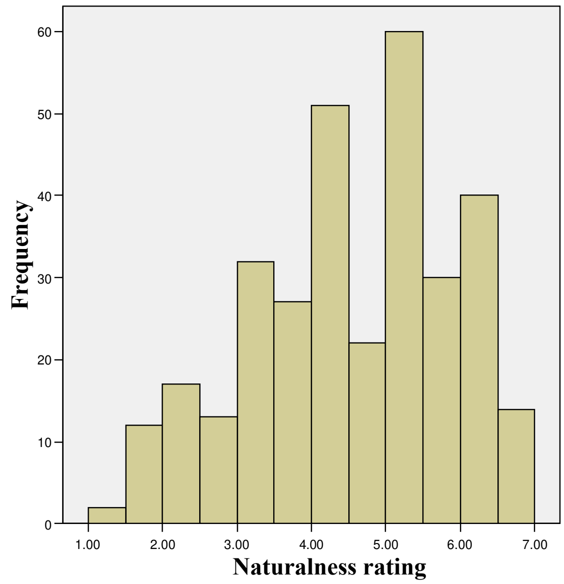


Figure 7.5: Distribution of naturalness ratings over the random utterance set (320 utterances). Ratings are averaged over all judges.

than the rule-based utterances, although Table 6.6 in Chapter 6 showed that some traits are perceived as more natural than others. An independent sample t-test shows that this difference is marginally significant ($p = .075$, two-tailed), which could result from inconsistent random generation decisions. While this difference reveals a weakness of stochastic approaches, a full evaluation needs to take into account the statistical selection model, that may reduce inconsistencies by biasing its selection towards more natural utterances.

Naturalness ratings can help identify what generation parameters are responsible for unnatural utterances. Table 7.4 shows the correlations between generation decisions and average naturalness ratings of the random utterances. Results show that negative content is perceived as unnatural ($r = -.32$), which might be due to the nature of the communicative goal, as recommendations rarely contain negative content. Negations ($r = -.27$), strict repetitions ($r = -.22$) and pronouns ($r = -.18$) also affect naturalness. On the other hand, positive content ($r = .32$) and in-group markers ($r = .24$) are perceived as natural, as well as more frequent words ($r = .21$). Interestingly, stuttering also increases naturalness ($r = .14$).

This analysis provides useful information for improving PERSONAGE's base generator. However, it is not clear whether the lack of naturalness results from im-

Generation decisions	r_{nat}
Content planning - content polarity	0.32**
Content planning - positive content	0.26**
Content planning - polarisation	0.26**
Pragmatic marker - in-group marker	0.24**
Lexical choice - lexicon frequency	0.21**
Aggregation - justify - since sn	0.17*
Pragmatic marker - in-group marker: pal	0.16*
Pragmatic marker - stuttering	0.14
Aggregation - infer - merge	0.12*
Aggregation - justify - since ns	0.12*
Pragmatic marker - tag question: okay	0.11*
Content planning - syntactic complexity	0.10
Content planning - concessions	-0.10
Pragmatic marker - softener: rather	-0.10
Aggregation - justify - with ns	-0.11
Pragmatic marker - filled pause: err	-0.11
Aggregation - concede - even if ns	-0.11*
Pragmatic marker - subject implicitness	-0.13*
Pragmatic marker - tag question: you see	-0.13*
Aggregation - restate - merge with comma	-0.13*
Aggregation - concede - but/though ns	-0.13*
Aggregation - infer - period	-0.14*
Pragmatic marker - softener: somewhat	-0.14*
Pragmatic marker - pronominalisation: demonstrative	-0.14*
Aggregation - justify - with ns	-0.14
Aggregation - restate - conjunction with comma	-0.14*
Pragmatic marker - softener: quite	-0.15**
Pragmatic marker - softener: sort of	-0.17**
Aggregation - concede - even if ns	-0.18
Content planning - repeated neutral content	-0.18**
Pragmatic marker - pronominalisation	-0.18**
Content planning - repetitions	-0.22**
Pragmatic marker - negation	-0.27**
Content planning - negative content	-0.32**

Table 7.4: Correlations between generation decisions and average naturalness ratings of the random utterances, at the $p < .1$ level (* = $p < .05$, ** = $p < .01$). Generation parameter names are prefixed with their component in the NLG architecture.

plementation issues that could be resolved, or from the personality markers themselves. Personality traits are normally distributed in the human population, thus rare linguistic markers of extreme traits—e.g. strict repetitions—may be likely to be interpreted as unnatural, especially in the absence of prosodic cues.

7.2 Feature analysis

This section explores the utility of a range of different utterance features, from the ones we can directly control—e.g. generation decision features—to ones that are only useful in a post-filtering phase. In order to evaluate the usefulness of individual features, we present an analysis of the correlation between the judges' ratings and (1) generation decision features, (2) content-analysis features and (3) n-gram frequency counts. These features are motivated by previous psychological findings about correlations between measurable linguistic factors and personality traits. See Chapter 4. Content-analysis categories were also used for recognising personality from general domain data in Chapter 3. These feature sets characterise different levels of utterance representation, from the semantic/pragmatic level (e.g. content planning generation decisions) to the surface realisation (e.g. frequency count of the word *bad*). Higher level features capture information about the generation process, whereas surface features capture emergent properties of the utterance, regardless of the underlying causes.

While some learning algorithms can learn from a large feature set, performance is generally increased when removing features uncorrelated with the target variable. The following subsections therefore analyse the correlation between each feature and the Big Five scores, based on the average ratings of the random utterances.[5]

7.2.1 Generation decisions

As PERSONAGE's generation decisions are motivated by psycholinguistic studies, we hypothesised that they affect the perception of the personality conveyed by an utterance. However, a correlational analysis shows that some generation decisions have a higher impact than others, and that some of the literature's findings summarised in Chapter 4 do not carry over to our domain. In order to factor out internal generation constraints that prevent PERSONAGE from always satisfying its input parameter values, the generation decision feature values used in this chapter are the actual decisions that were taken in each utterance rather than input parameter values, e.g. the CONCESSIONS feature represents the actual number of CONCEDE rhetorical relations produced, not the input selection probability of the aggregation operation. Generation decision features are labelled with the generation parameter's name prefixed with its component in the NLG architecture. Some aggregation

[5]The utterances generated from predefined parameter settings in Chapter 6 are not used, as their feature values are highly dependent within each utterance set. This dependence prevents learning algorithms from identifying the features responsible for the correlation with the ratings.

features count the occurrences of an operation with a specific rhetorical relation (e.g. AGGREGATION - INFER - PERIOD), while others count aggregation operations used with multiple relations (e.g. AGGREGATION - CONJUNCTION), or occurrences of a rhetorical relation regardless of the operation used (e.g. AGGREGATION - INFER). While our experiments test multiple hypotheses on the same data, we do not adjust the reported significance levels throughout this chapter (e.g. Bonferroni correction), in order to make sure that all significant relations are identified, at the risk of inflating the significance of some of them.

Generation decision features	r_{extra}	**Pred**
PRAGMATIC MARKER - EXCLAMATION	0.34**	yes
AGGREGATION - INFER	0.21**	no
CONTENT PLANNING - VERBOSITY	0.19**	yes
CONTENT PLANNING - REQUEST CONFIRMATION: YOU WANT TO KNOW	0.16**	yes
CONTENT PLANNING - REQUEST CONFIRMATION: DID YOU SAY	0.16*	yes
AGGREGATION - JUSTIFY - SINCE NS	0.16	no
AGGREGATION - CONJUNCTION	0.16**	yes
LEXICAL CHOICE - LEXICON FREQUENCY	0.15*	yes
PRAGMATIC MARKER - NEAR EXPLETIVES	0.15	yes
CONTENT PLANNING - SYNTACTIC COMPLEXITY	0.15**	opp
PRAGMATIC MARKER - EMPHASISER: REALLY	0.14*	yes
CONTENT PLANNING - REQUEST CONFIRMATION	0.14*	yes
AGGREGATION - RESTATE - CONJUNCTION WITH COMMA	0.13*	yes
AGGREGATION - INFER - PERIOD	0.13*	opp
CONTENT PLANNING - RESTATEMENTS	0.13*	yes
AGGREGATION - RESTATE	0.12*	yes
AGGREGATION - INFER - CONJUNCTION	0.12*	yes
CONTENT PLANNING - REPEATED POSITIVE CONTENT	0.12*	yes
AGGREGATION - PERIOD	0.12*	yes
CONTENT PLANNING - TEMPLATE POLARITY	0.12*	yes
CONTENT PLANNING - REPETITION POLARITY	0.12*	yes
AGGREGATION - INFER - MERGE	0.11	no
PRAGMATIC MARKER - SOFTENER: KIND OF	-0.10	yes
PRAGMATIC MARKER - SOFTENER: RATHER	-0.11*	yes
PRAGMATIC MARKER - SOFTENER: LIKE	-0.11*	yes
CONTENT PLANNING - INITIAL REJECTION	-0.18*	no
PRAGMATIC MARKER - FILLED PAUSE: ERR	-0.23**	yes

Table 7.5: Correlations between generation decision features and average extraversion ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$). The *Pred* column indicates whether the relation was predicted by the psychology findings reviewed in Chapter 4 (*opp* = predicted opposite relation).[6]

Table 7.5 shows that exclamation marks are the strongest indicators of extraversion, with a correlation of .34 with the average ratings. As suggested by the

[6]Extraversion ratings were collected over two experiments, with different generation parameters. As both resulting datasets are merged together, generation decisions that were implemented in both experiments produce higher significance levels.

literature, verbosity is also associated with extraversion, however the use of the INFER rhetorical relation—joining propositions together without emphasis—produces a higher association, suggesting that extraverts do not put pieces of information into perspective.[7] Explicit confirmations are also associated with extraversion, as well as the use of conjunctions, frequent lexical items, near-expletives (e.g. *darn*), the adverb *really*, restatements, and more positive claims. Negative correlations at the bottom of Table 7.5 indicate markers of introversion. The filled pause *err* is the strongest indicator of introversion, with a correlation of $-.23$. Introverts are also perceived as producing initial rejections, as well as hedges such as *like* and *rather*.

Generation decision features	r_{emot}	**Pred**
LEXICAL CHOICE - LEXICON WORD LENGTH	0.25**	no
PRAGMATIC MARKER - IN-GROUP MARKER: PAL	0.22**	no
PRAGMATIC MARKER - IN-GROUP MARKER	0.20**	no
AGGREGATION - JUSTIFY - SINCE NS	0.16*	no
PRAGMATIC MARKER - ACKNOWLEDGMENT: YEAH	0.15	opp
CONTENT PLANNING - CONTENT POLARITY	0.14	yes
AGGREGATION - WITH	-0.13	yes
CONTENT PLANNING - RESTATEMENTS	-0.15	no
AGGREGATION - INFER - WITH NS	-0.15	yes
AGGREGATION - INFER - ALSO	-0.16	opp
PRAGMATIC MARKER - ACKNOWLEDGMENT: OK	-0.16*	yes
AGGREGATION - MERGE	-0.16*	yes
AGGREGATION - CONCEDE - ALTHOUGH NS	-0.17*	no
CONTENT PLANNING - REPETITIONS	-0.18*	yes
PRAGMATIC MARKER - EXPLETIVES	-0.18*	yes
AGGREGATION - RESTATE - MERGE WITH COMMA	-0.19*	yes
PRAGMATIC MARKER - TAG QUESTION	-0.19*	yes
CONTENT PLANNING - SYNTACTIC COMPLEXITY	-0.19*	no
CONTENT PLANNING - NEGATIVE CONTENT	-0.20*	yes
PRAGMATIC MARKER - TAG QUESTION: ALRIGHT	-0.21**	yes
PRAGMATIC MARKER - EXPLETIVES: DAMN	-0.21**	yes
PRAGMATIC MARKER - FILLED PAUSE: ERR	-0.22**	yes
AGGREGATION - RESTATE	-0.23**	yes
AGGREGATION - JUSTIFY - SO SN	-0.25*	opp
CONTENT PLANNING - REPEATED NEGATIVE CONTENT	-0.26**	yes
LEXICAL CHOICE - LEXICON FREQUENCY	-0.28**	yes

Table 7.6: Correlations between generation decision features and average emotional stability ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$). The *Pred* column indicates whether the relation was predicted by the psychology findings reviewed in Chapter 4 (*opp* = predicted opposite relation).

As far as emotional stability is concerned, Table 7.6 shows that neuroticism is associated with the use of short, frequent words ($r = -.28$). Neurotics are also

[7]To improve readability throughout this chapter, the perception of the judges regarding a personality type is referred to as a characteristic of individuals that possess that personality trait, e.g. '*extraverts use more exclamation marks*'.

perceived as using the discourse connective *so* to express justifications, whereas *since* is associated with stable speakers. Interestingly, in-group markers indicate stability as well (especially *pal*), whereas filled pauses (i.e. *err*) and repetitions indicate neuroticism. As suggested by the literature, negative content and swear words are also associated with a lack of stability, with a stronger association for the expletive *damn* ($r = -.21$).

Table 7.7 shows that agreeableness is the trait presenting the highest correlation with language generation decisions. Polarity is the most important indicator of agreeableness, with correlations of .49 for the CONTENT POLARITY parameter and .16 for the polarity of repeated propositions, i.e. disagreeable speakers select and emphasise negative content. Also, although agreeableness is associated with positive content, it is not marked by positively-connotated claim templates. The second most important marker of disagreeableness is the use of the PERIOD operation for contrasting propositions ($r = -.41$), which can be perceived as a long, unfilled pause. On the other hand, exclamation marks and long words are perceived as agreeable. As suggested by the literature on politeness [Brown and Levinson, 1987], in-group markers also project agreeableness ($r = .33$), especially *pal*, as well as the hedging expression *it seems to me that* ($r = .15$). Interestingly, *it seems that* produces the opposite effect ($r = -.20$), suggesting that empathy with the hearer is an important aspect of agreeableness. While the data confirms hypothesised markers—such as swear words ($r = -.16$), many unforeseen markers of disagreeableness emerge: concessions, repetitions of the proposition's object (RESTATE - MERGE WITH COMMA), filled pauses, verbosity, demonstrative referring expressions and tag questions. These new findings suggest that personality models can benefit from a domain-specific data-driven analysis.

Correlations for conscientiousness in Table 7.8 are not as high as for previous traits, suggesting that this dimension is more difficult to model in our domain. Negative content is strongly associated with a lack of conscientiousness ($r = -.31$), as well as concessions, repetitions, tag questions ($r = -.22$), the *so* cue word for expressing justifications ($r = -.19$), filled pauses ($r = -.18$), and the use of frequent words ($r = -.17$). Concerning the positive end of the scale, in-group markers ($r = .23$), polarity ($r = .21$), the *since* cue word and the MERGE operation all indicate conscientiousness. Longer words are also perceived as conscientious,

Generation decision features	r_{agree}	**Pred**
CONTENT PLANNING - CONTENT POLARITY	0.49**	yes
CONTENT PLANNING - POSITIVE CONTENT	0.37**	yes
PRAGMATIC MARKER - IN-GROUP MARKER	0.33**	yes
CONTENT PLANNING - POLARISATION	0.25**	no
PRAGMATIC MARKER - IN-GROUP MARKER: PAL	0.24**	yes
LEXICAL CHOICE - LEXICON WORD LENGTH	0.21**	yes
PRAGMATIC MARKER - IN-GROUP MARKER: MATE	0.17*	yes
PRAGMATIC MARKER - EXCLAMATION	0.17*	no
CONTENT PLANNING - REPETITION POLARITY	0.16*	yes
PRAGMATIC MARKER - SOFTENER: IT SEEMS TO ME THAT	0.15	yes
PRAGMATIC MARKER - TAG QUESTION: ALRIGHT	-0.14	opp
PRAGMATIC MARKER - SOFTENER: KIND OF	-0.14	opp
CONTENT PLANNING - POSITIVE CONTENT FIRST	-0.15	opp
PRAGMATIC MARKER - PRONOMINALISATION	-0.15	no
AGGREGATION - INFER	-0.15	no
AGGREGATION - JUSTIFY - PERIOD	-0.15	yes
CONTENT PLANNING - COMPETENCE MITIGATION: COME ON	-0.16	yes
PRAGMATIC MARKER - SOFTENER: QUITE	-0.16*	opp
PRAGMATIC MARKER - EXPLETIVES	-0.16*	yes
AGGREGATION - WITH	-0.16*	no
AGGREGATION - RESTATE - MERGE WITH COMMA	-0.17*	no
PRAGMATIC MARKER - TAG QUESTION: YOU SEE	-0.17*	opp
PRAGMATIC MARKER - EXPLETIVES: DAMN	-0.18*	yes
PRAGMATIC MARKER - SOFTENER: SOMEWHAT	-0.18*	opp
PRAGMATIC MARKER - SUBJECT IMPLICITNESS	-0.18*	yes
AGGREGATION - PERIOD	-0.18*	yes
AGGREGATION - ALTHOUGH	-0.18*	no
AGGREGATION - CONCEDE - ALTHOUGH NS	-0.18*	no
AGGREGATION - INFER - WITH NS	-0.18*	no
AGGREGATION - EVEN IF	-0.19*	no
PRAGMATIC MARKER - TAG QUESTION	-0.19*	no
PRAGMATIC MARKER - PRONOMINALISATION: DEMONSTRATIVE	-0.20*	no
PRAGMATIC MARKER - SOFTENER: IT SEEMS THAT	-0.20*	opp
AGGREGATION - CONCEDE - EVEN IF NS	-0.21**	no
AGGREGATION - CONTRAST - PERIOD	-0.25**	yes
CONTENT PLANNING - VERBOSITY	-0.28**	no
AGGREGATION - JUSTIFY - PERIOD	-0.28*	yes
PRAGMATIC MARKER - FILLED PAUSE: ERR	-0.28**	no
CONTENT PLANNING - CONCESSIONS	-0.29**	no
AGGREGATION - CONCEDE	-0.29**	no
CONTENT PLANNING - REPEATED NEGATIVE CONTENT	-0.32**	yes
AGGREGATION - CONTRAST - PERIOD	-0.41**	yes
CONTENT PLANNING - NEGATIVE CONTENT	-0.53**	yes

Table 7.7: Correlations between generation decision features and average agreeableness ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$). The *Pred* column indicates whether the relation was predicted by the psychology findings reviewed in Chapter 4 (*opp* = predicted opposite relation).

possibly because they tend to be less frequent, thus indicating that the speaker has a larger vocabulary.

Openness to experience is the hardest trait to model in our domain, as cor-

Generation decision features	r_{consc}	**Pred**
PRAGMATIC MARKER - IN-GROUP MARKER	0.23**	opp
CONTENT PLANNING - CONTENT POLARITY	0.21**	yes
AGGREGATION - JUSTIFY - SINCE NS	0.21	yes
PRAGMATIC MARKER - ACKNOWLEDGMENT: YEAH	0.19*	opp
LEXICAL CHOICE - LEXICON WORD LENGTH	0.18*	yes
AGGREGATION - INFER - MERGE	0.17	no
CONTENT PLANNING - REQUEST CONFIRMATION: LET'S SEE WHAT	0.16*	yes
CONTENT PLANNING - NEUTRAL CONTENT	0.15	no
AGGREGATION - JUSTIFY	-0.13	yes
CONTENT PLANNING - INITIAL REJECTION: I'M NOT SURE	-0.14	yes
CONTENT PLANNING - INITIAL REJECTION	-0.15	yes
PRAGMATIC MARKER - TAG QUESTION: ALRIGHT	-0.15	no
AGGREGATION - ALTHOUGH	-0.15	opp
AGGREGATION - RESTATE - MERGE WITH COMMA	-0.15	no
PRAGMATIC MARKER - SOFTENER: QUITE	-0.16*	opp
PRAGMATIC MARKER - PRONOMINALISATION	-0.16*	no
AGGREGATION - RESTATE	-0.16*	no
AGGREGATION - CONCEDE - BUT/THOUGH NS	-0.17*	yes
LEXICAL CHOICE - LEXICON FREQUENCY	-0.17*	yes
PRAGMATIC MARKER - FILLED PAUSE: ERR	-0.18*	yes
AGGREGATION - JUSTIFY - SO SN	-0.19	yes
CONTENT PLANNING - REPEATED NEGATIVE CONTENT	-0.19*	yes
AGGREGATION - CONTRAST - PERIOD	-0.21*	no
PRAGMATIC MARKER - TAG QUESTION	-0.22**	no
CONTENT PLANNING - REPETITIONS	-0.22**	yes
CONTENT PLANNING - CONCESSIONS	-0.22**	opp
AGGREGATION - CONCEDE	-0.22**	opp
AGGREGATION - CONCEDE - ALTHOUGH NS	-0.27*	opp
CONTENT PLANNING - NEGATIVE CONTENT	-0.31**	yes

Table 7.8: Correlations between generation decision features and average conscientiousness ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$). The *Pred* column indicates whether the relation was predicted by the psychology findings reviewed in Chapter 4 (*opp* = predicted opposite relation).

relations in Table 7.9 do not exceed .23. A possible cause is that it is the most controversial of the Big Five traits [Goldberg, 1990]: openness to experience is the dimension explaining the smallest amount of variance in factor analyses. Also, some researchers have given a different name to this dimension, such as 'culture' or 'intellect'. Results show that openness to experience is partly expressed using similar cues as conscientiousness, i.e. positive content and in-group markers, together with the avoidance of filled pauses and tag questions. Interestingly, exclamation marks and explicit confirmations also indicate openness ($r = .17$), whereas the use of *with* for aggregation and the back-channel *ok* both indicate narrow-mindedness ($r = -.20$ and $-.16$).

The correlational analysis presented here provides insight into what genera-

Generation decision features	r_{open}	Pred
CONTENT PLANNING - TEMPLATE POLARITY	0.23**	yes
PRAGMATIC MARKER - IN-GROUP MARKER	0.22**	no
CONTENT PLANNING - REQUEST CONFIRMATION: LET'S SEE WHAT	0.19*	yes
PRAGMATIC MARKER - EXCLAMATION	0.17*	no
CONTENT PLANNING - CONTENT POLARITY	0.17*	yes
CONTENT PLANNING - REQUEST CONFIRMATION	0.17*	yes
CONTENT PLANNING - REPETITION POLARITY	0.16*	yes
PRAGMATIC MARKER - IN-GROUP MARKER: PAL	0.15	no
AGGREGATION - PERIOD	0.14	no
PRAGMATIC MARKER - SOFTENER: I THINK THAT	0.13	yes
AGGREGATION - JUSTIFY - SO SN	-0.13	yes
PRAGMATIC MARKER - TAG QUESTION: OKAY	-0.13	opp
PRAGMATIC MARKER - ACKNOWLEDGMENT: OK	-0.16*	no
PRAGMATIC MARKER - TAG QUESTION: YOU SEE	-0.17*	opp
PRAGMATIC MARKER - TAG QUESTION	-0.18*	opp
CONTENT PLANNING - REPEATED NEGATIVE CONTENT	-0.19*	yes
CONTENT PLANNING - NEGATIVE CONTENT	-0.19*	yes
AGGREGATION - INFER - WITH NS	-0.20*	opp
PRAGMATIC MARKER - FILLED PAUSE: ERR	-0.20*	yes

Table 7.9: Correlations between generation decision features and average openness to experience ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$). The *Pred* column indicates whether the relation was predicted by the psychology findings reviewed in Chapter 4 (*opp* = predicted opposite relation).

tion parameters are helping the judges to discriminate between various traits. The knowledge of strong markers of personality is useful for controlling the generation process. Also, these correlations offer an evaluation of the hypotheses made in Chapter 4, i.e. domain-specific findings can carry over to project personality in PERSONAGE's domain. Interestingly, we find that many new markers emerge, while some results contradict our hypotheses (i.e. indicated by *opp* in the *Pred* columns). Future work could thus enhance PERSONAGE-RB's rule-based approach based on the correlations presented here, by taking domain-specific information into account to refine the predefined parameter settings derived from psychological studies.

7.2.2 Content-analysis features

Content-analysis tools are handcrafted dictionaries that are typically used for studying psycholinguistic properties of texts or utterances, in order to provide information about the author or speaker. They have been used for identifying the relation between language and personality for different genres [Pennebaker and King, 1999, Gill and Oberlander, 2002, Mehl et al., 2006], as well as in the personality recognition models in Chapter 3.

The first content-analysis utility used in this thesis is the Linguistic Inquiry and Word Count (LIWC) tool [Pennebaker et al., 2001], which consists of 88 word categories including both syntactic (e.g. ratio of pronouns) and semantic information (e.g. positive emotion words), which were validated by expert judges. LIWC features are illustrated in Tables 3.4 and 3.5 in Chapter 3. LIWC features—apart from the word count and the number of words per sentence—are all expressed in terms of the percentage of words in the utterance belonging to the corresponding word category. Pennebaker and King [1999] and Mehl et al. [2006] previously found significant correlations between these features and each of the Big Five personality traits.

The second utility is the MRC psycholinguistic database [Coltheart, 1981], which contains statistics for over 150,000 words, such as estimates of the age of acquisition, concreteness, frequency of use and familiarity. Table 3.4 in Chapter 3 shows examples of MRC scales. Each MRC feature is computed by averaging the feature value of all content words in the utterance, i.e. lexemes in the Deep Syntactic Structure (DSyntS). Part-of-speech tags are extracted from the utterance's DSyntS to discriminate between homonym entries.

While a correlational analysis of these features on general conversational data was presented in Section 3.2.3 in Chapter 3, the current section focuses on personality correlates in PERSONAGE's information presentation domain.

As content-analysis features capture both syntactic and semantic information, some are equivalent to generation decision features, such as the number of exclamation marks.[8] However, content-analysis features can also capture multiple generation decisions. For example, the word count feature is strongly related to the VERBOSITY parameter as well as verbose aggregation operations (e.g. CONJUNCTION), and the number of positive emotion words is influenced by multiple polarity-related parameters. Thus, the following analysis focuses primarily on previously unidentified markers.

As in the previous section, Table 7.10 shows that exclamation marks are the strongest indicator of extraversion. The use of full stops indicates introversion ($r = -.25$), which is likely to result from both hesitancy markers (i.e. *err...*) and

[8]The LIWC exclamation mark feature differs slightly from the generation decision as the former counts the total number of exclamation marks divided by the utterance length, whereas the latter counts whether an exclamation mark was inserted by the generator or not.

Type	Features	r_{extra}
LIWC	Exclamation marks	0.33**
LIWC	Word count (WC)	0.19**
LIWC	Achievement	0.13*
LIWC	Words per sentence (WPS)	0.13*
LIWC	Optimism	0.12*
MRC	Familiarity	0.12*
LIWC	Dictionary	0.11
LIWC	Physical states	0.10
LIWC	Time	-0.10
LIWC	Negative emotion	-0.10
LIWC	Tentative	-0.13*
LIWC	Assents	-0.15**
LIWC	Negations	-0.15**
LIWC	All punctuation	-0.21**
LIWC	Period	-0.25**

Table 7.10: Correlations between LIWC and MRC features and average extraversion ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$).

the PERIOD aggregation operation. While the previous section does not show any correlation between extraversion and the NEGATION parameter (see Table 7.5), the corresponding LIWC feature does capture the association, possibly due to a preference for negative claim templates (e.g. *this restaurant is not as bad as the others*). Additionally, the negative correlation with tentativeness (e.g. *seem*, *guess*, *kind of*) reflects the hypothesised preferences of introverts for hedging and understatement, whereas related generation parameters do not emerge as strongly in Table 7.5. Extraversion is marked by longer sentences, suggesting the avoidance of the PERIOD and MERGE aggregation operations. Interestingly, although Table 7.5 indicates that extraverts do not select more positive attributes at the content level, they use more words related to achievement and optimism (e.g. *best*, *excellent*). Whereas LIWC features provide many significant correlations for extraversion, MRC features only reveal that extraverts are perceived as using more familiar words.

Results for emotional stability in Table 7.11 show that references to other people and communication are the main indicators of stability, as words related to social processes (e.g. *pal*, *suggest*, second person pronouns), friends, and family are all positively correlated with that trait (from $r = .16$ to $r = .26$). This finding was already hinted by the association with in-group markers shown in the previous section (see Table 7.6). As suggested by the literature, neuroticism is strongly associated with negative affect (e.g *bad*, *awful*) and swearing.

Similarly to emotional stability, Table 7.12 shows that agreeableness is asso-

Type	Features	r_{emot}
LIWC	Social processes	0.26**
LIWC	Friends	0.24**
LIWC	Prepositions	0.16*
LIWC	Family	0.16*
LIWC	People	0.13
LIWC	Question marks	-0.13
MRC	Age of acquisition (AOA)	-0.14
LIWC	Anger	-0.15
LIWC	All punctuation	-0.23**
LIWC	Affective processes	-0.23**
LIWC	Swear words	-0.25**
LIWC	Negative emotion	-0.25**

Table 7.11: Correlations between LIWC and MRC features and average emotional stability ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$).

Type	Features	r_{agree}
LIWC	Friends	0.34**
LIWC	Social processes	0.27**
LIWC	Self	0.23**
LIWC	I	0.21**
LIWC	Exclamation marks	0.20*
LIWC	Pronouns	0.19*
LIWC	Discrepancy	0.16*
LIWC	Causation	0.16*
LIWC	Positive feeling	0.15
LIWC	People	0.15
LIWC	Sexuality	0.15
LIWC	Communication	0.14
LIWC	You	0.14
LIWC	We	0.13
LIWC	Cognitive processes	0.13
LIWC	Affective processes	-0.15
LIWC	Tentative	-0.18*
LIWC	Six letters	-0.18*
LIWC	Word count (WC)	-0.19*
MRC	Age of acquisition (AOA)	-0.22**
LIWC	Anger	-0.22**
LIWC	Swear words	-0.22**
LIWC	Negative emotion	-0.33**

Table 7.12: Correlations between LIWC and MRC features and average agreeableness ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$).

ciated with the use of words related to social processes ($r = .27$) and friends ($r = .34$), including in-group markers. Interestingly, agreeable utterances contain more self-references (e.g. first person pronouns), which was not captured by generation decision features. On the other hand, LIWC features fail to capture the strong association between agreeableness and positive content. Agreeableness is also associated with pronouns and words related to discrepancies (e.g. *would*, *if*).

As hypothesised, disagreeable utterances contain more negative affect and swearing. However, they also contain longer words with a high age of acquisition, as well as expressions of tentativeness, which were not captured by generation decisions.

Type	Features	r_{consc}
LIWC	FRIENDS	0.21**
LIWC	SOCIAL PROCESSES	0.20*
MRC	CONCRETENESS	0.16*
MRC	IMAGERY	0.13
LIWC	OTHER REFERENCES	0.14
LIWC	NEGATIONS	-0.13
LIWC	CERTAINTY	-0.13
LIWC	SWEAR WORDS	-0.15
LIWC	ANGER	-0.16*
LIWC	QUESTION MARKS	-0.19*
LIWC	AFFECTIVE PROCESSES	-0.26**
LIWC	NEGATIVE EMOTION	-0.30**

Table 7.13: Correlations between LIWC and MRC features and average conscientiousness ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$).

LIWC features in Table 7.13 show that conscientiousness is associated with friends and social processes ($r = .20$), whereas unconscientious utterances contain more negative affect ($r = -.30$) and questions ($r = -.19$). Also, MRC features reveal that conscientiousness can be conveyed by using more concrete words (e.g. *restaurant*), as well as words conveying imagery (e.g. *table*, *car*).

Concerning openness to experience, Table 7.14 shows that concreteness and imagery are the strongest markers of that trait ($r = .24$). Open speakers also refer more to social processes and acquaintances, as well as to the exact price rather than subjective modifiers (e.g. $44 vs. *cheap*, *expensive*). As shown in the last section, narrow-mindedness is expressed through negative affect and tag questions. However, LIWC features do not capture the filled pause *err*, as word categories sometimes fail to include spoken onomatopoeic terms. Worlds related to assent (e.g. *ok*, *yeah*, *alright*) also indicate narrow-mindedness, which result from acknowledgment generation parameters as well as the *alright?* tag question.

As content-analysis categories were handcrafted by psychologists, they are general enough to detect the speaker's psychological dispositions from a relatively short amount of text. But could some categories be *too* general for the generation task, i.e. aggregating together unrelated cues in our domain? The next section tries to answer this question by analysing correlates of personality at the word level.

Type	Features	r_{open}
MRC	CONCRETENESS	0.24**
MRC	IMAGERY	0.23**
LIWC	EXCLAMATION MARKS	0.20**
LIWC	SOCIAL PROCESSES	0.17*
LIWC	FRIENDS	0.17*
LIWC	UP	0.17*
LIWC	NUMBERS	0.17*
LIWC	PREPOSITIONS	0.16*
MRC	NUMBER OF PHONEMES	0.15
LIWC	POSITIVE FEELING	0.15
LIWC	SEXUALITY	0.15
LIWC	OPTIMISM	0.14
LIWC	AFFECTIVE PROCESSES	-0.17*
LIWC	ASSENTS	-0.20*
LIWC	NEGATIVE EMOTION	-0.20*
LIWC	QUESTION MARKS	-0.23**

Table 7.14: Correlations between LIWC and MRC features and the average openness to experience rating at the $p < .1$ level (* = $p < .05$, ** = $p < .01$).

7.2.3 N-gram features

The relation between n-gram frequency counts in emails and the personality of the author has been studied by Gill and Oberlander [2002], for both extraversion and emotional stability. As individual words are the unit of analysis, n-gram studies are highly domain-dependent. Thus, an analysis of PERSONAGE's output is likely to reveal new personality markers. While a typical weakness of n-gram studies is the sparsity of the n-gram counts—resulting in poor real-world estimates, PERSONAGE's domain offers the advantage of being relatively constrained, with a vocabulary of around 270 words without database-specific terms. In order to reduce the sparsity of the feature space, restaurant names, cuisine types and numbers were replaced by generic labels in the surface realisation (in upper case). Table 7.15 shows the unigrams, bigrams and trigrams that are the most highly correlated with the average extraversion ratings.[9] To keep this analysis within reasonable length, only n-grams associated with extraversion are presented.

Although n-gram features do not incorporate any psycholinguistic knowledge, they can capture previously identified markers, such as exclamation marks ($r = .34$), the positive template '*X is the best restaurant of his kind*' ($r = .20$), and specific request confirmations ('*you want to know more about RESTAURANT?*', $r = .14$).

[9]Redundant n-grams were discarded, i.e. if they co-occur exactly with one of the n-grams on the list.

Features	r_{extra}
!	0.34**
pal#!	0.25**
of#its#kind	0.20**
nice#food#.	0.18**
,#it	0.17**
is#just#located	0.17*
best#restaurant#of	0.15**
best#place	0.15**
yeah#,#oh	0.15
its	0.15**
since#it#s	0.15
it#provides#nice	0.15
about#RESTAURANT#?	0.14*
offers#mediocre#ambience	0.14
mmhm#...#basically	0.14
appreciate#RESTAURANT#since	0.14
s#just#bloody	0.14
place#,#so	0.14
consider#,#also	0.14
t#quite#nasty	0.14
dollars#,#its	0.14
although#RESTAURANT#provides	0.14
ok#,#it	0.14
,#it#s	0.14*
them#!	0.14
best#place#of	0.14*
RESTAURANT#is#...	0.14
!#RESTAURANT	0.14
is#darn#inexpensive	0.13
it	0.13*
it#s#a	0.13*
oh#gosh	0.13
s#an#CUISINE	0.13
the#best#restaurant	0.13*
offers#poor#ambience	0.13
adore#it	0.13
RESTAURANT#!	0.13
outstanding	0.13*
doesn#t#offer	0.13
place#of#its	0.13*
since	0.13*
restaurant#,#which	0.13*
s#a#CUISINE	0.13*
this#restaurant#,	0.13*
,#it#features	0.12*
good#food	0.12*
you#know#!	0.12*
actually#,#basically	0.12*
NUMBER#dollars#!	0.12*
dollars#!	0.12*
food#.#it	0.12*
a#CUISINE#restaurant	0.12*
quite#friendly#waiters	0.12*
restaurants#.#RESTAURANT	0.12*
has	0.12*
its#kind#since	0.12*
mean#,#it	0.11*
restaurant#,	0.11*
friendly	0.11*
quite#friendly	0.11*
just#located#in	0.11
best	0.11
it#s#really	0.11
and	0.11
although#its#price	0.11
cheap#,#it	0.11
and#this#restaurant	0.11
I#would#recommend	0.11

Features	r_{extra}
satisfactory#food	-0.11
quite#nice	-0.11
nice#service#,	-0.11*
t#expensive#,	-0.11*
RESTAURANT#features#acceptable	-0.11*
just#nice	-0.11*
ok#,	-0.11*
that#RESTAURANT#,	-0.11*
bad#waiters	-0.12*
is#rather#nice	-0.12*
of#dainty#food	-0.12*
kind#of#nice	-0.12*
is#rather	-0.12*
t#kind#of	-0.12*
quite#satisfying	-0.12*
sort#of	-0.12*
t#somewhat	-0.12*
rather#cheap#.	-0.12*
I	-0.13*
also#its#price	-0.13*
place#with#nice	-0.13*
okay#?	-0.13*
not#sure#.	-0.13
it#just#has	-0.13*
RESTAURANT	-0.13*
somewhat#good#.	-0.13
I#guess	-0.14
well#,#yeah	-0.14
it#offers#sort	-0.14
dainty#food#.	-0.14
appreciate#RESTAURANT#,	-0.14
?#although#the	-0.14
just#acceptable#.	-0.14
would#approve#them	-0.14
.#I#suppose	-0.14
somewhat#inexpensive#.	-0.14
is#somewhat#passable	-0.14
ob-ob-obviously#,	-0.14
,#ba-ba-basically#,	-0.14
would#appreciate#it	-0.14
sort#of#wrong	-0.14
quite#nice#service	-0.14
bad	-0.15**
err#...#actually	-0.15**
pal#.	-0.15
it#offers#like	-0.15
advise#them#.	-0.15
provides#rather#adequate	-0.15
LOCATION#with#passable	-0.15
might#advise#them	-0.15
?#I#guess	-0.15
and#quite#bad	-0.15
low-cost#with#kind	-0.15
food#,#basically	-0.15
...#actually#,	-0.16**
offers#like#,	-0.16
but#RESTAURANT#is	-0.16*
has#satisfactory#food	-0.17*
.#err#...	-0.17**
because#RESTAURANT#is	-0.17*
it#,#okay	-0.18*
sort#of#alright	-0.18*
see#?#although	-0.18*
alright	-0.19**
of#bad#waiters	-0.21**
...	-0.22**
of#alright	-0.22**
err#...	-0.24**

Table 7.15: Correlations between n-gram features and average extraversion ratings at the $p < .1$ level (* = $p < .05$, ** = $p < .01$).

New combinations also emerge, such as the in-group marker *pal* followed by an exclamation mark ($r = .25$), the adjective *nice* for describing food ($r = .18$), the phrase *'X is just located in Y'* ($r = .17$) and the hedging expression *'you know!'* ($r = .12$).

New markers of introversion include the hedges *kind of* and *sort of* modifying the low polarity adjective *alright* ($r = -.22$) or *bad waiters* ($r = -.21$), as well as more formal verbs such as *offer*, *advise* and *provide*, especially when combined with softening hedges, e.g. *'I might advise X'* and *'Y provides rather adequate Z'* ($r = -.15$).

More generally, introvert n-grams tend to be related to hesitation, negative content and softening hedges, whereas extraversion is associated with exclamations, questions and social hedges.

7.3 Discussion and summary

The first part of this chapter shows that PERSONAGE can generate utterances covering a large range of each of the Big Five dimensions, while some utterances are even perceived as more extreme than those generated using the psychologically-informed rule-based approach in the previous chapter. While the judges agree significantly over the personality of the utterances, their level of agreement is weaker than on utterances generated using the rule-based approach. This difference is likely to result from the random utterances' inconsistent personality cues. These inconsistencies also explain why random utterances are perceived as slightly less natural (.21 decrease on a scale from 1 to 7), although utterances are still perceived as natural on average (4.38 out of 7). These results suggest that PERSONAGE could be used for data-driven generation over a continuous scale, if controlled accurately by personality models. Data-driven techniques for controlling the personality conveyed by the generator are presented and evaluated in the next two chapters.

The correlational analysis presented in the previous section reveals some advantages and weaknesses of each feature set. Generation decision features model deep phenomena with good accuracy, as they control high-level intentions that can result in multiple surface realisations. However, their number is constrained by the generator's implementation, thus the level of control is limited to the generation decisions that were expected—at development time—to have a salient influence on the projected personality.

Content-analysis features are useful for identifying general psychological aspects of the speaker that result from combinations of generation decisions, although word sense ambiguity can produce misleading results in narrow domains, e.g. the expletive '*oh God*' is interpreted as related to religion and metaphysical issues by the LIWC utility. Nevertheless, such tools are the result of a large effort to define suitable categories aiming at detecting psychological dispositions of the speaker, including personality. Thus, the inclusion of this expert knowledge is likely to improve the generality of data-driven models.

N-gram frequency counts are the only features derived from the data without any assumption, they can therefore reveal unexpected findings. However, they do not model long-distance dependencies; and their low frequency counts can result in unreliable correlation estimates. Additionally, the high-granularity of n-gram features makes them very sensitive to training data, and therefore unlikely to generalise to other domains (see Section 8.4). As personality traits are predictive of general patterns of behaviour over time and situations, they are not expected to predict single behaviours accurately [Daly and Bippus, 1998], such as the production of a specific sequence of words.

Furthermore, significant correlations of generation decision features with personality ratings validate the hypotheses made in Chapter 4 regarding the generalisation of personality markers to PERSONAGE's domain. For example, the VERBOSITY parameter is strongly associated with extraversion, while expletives are perceived as strongly disagreeable. However, some hypothesised markers do not generalise, e.g. a high CONTENT POLARITY value does not convey extraversion successfully, and the use of expletives does not correlate negatively with conscientiousness ratings. Interestingly, this analysis also reveals personality markers that were not previously identified by psychological studies, e.g. disagreeableness is marked by the use of concessions, the discourse connective *so* is associated with neuroticism, and in-group markers indicate openness to experience. Such results could be used for refining the hypothesised parameter settings presented in Chapter 4.

The correlational analysis presented in this chapter provides insight onto what utterance properties convey specific personality traits. While the analysis of annotated output data can be used to calibrate a rule-based generator, it can also be used directly to learn statistical models that control the generation process by combining relevant features together. Chapters 8 and 9 investigate two methods for learning such models.

Chapter 8

Data-driven Generation of Personality through Overgeneration

The previous chapter established that PERSONAGE's random generation can produce utterances normally distributed across the Big Five dimensions, covering a large range of each scale. This is a prerequisite sampling issue, that needs to be established before we can consider training for such a sample or selecting from this sample with the goal of hitting particular targets within the normal distribution. In this chapter, we examine whether we can generate personality using the *overgenerate and select* method—a stochastic data-driven generation paradigm that has been used in previous work for optimising utterance quality [Langkilde-Geary, 2002, Walker et al., 2002]. We implement this approach in the PERSONAGE-OS data-driven generator, which overgenerates using the PERSONAGE base generator and selects the output utterance with the desired personality using statistical models trained on user feedback. This is the first application of this method to the generation of continuous personality variation, as well as its first application for generating specific targets along a scale, rather than ranking for the purpose of selecting only highly ranked outputs. In chapter 9, we use the same set of random utterances to develop and test an alternate and novel data-driven generation method.

8.1 Methodology

The overgenerate and select approach consists of two phases: (1) an *overgeneration* phase in which a base generator produces many candidate utterances expressing a target communicative goal, and (2) a *selection* phase in which a model selects the utterance maximising an objective function. Previous work has maximised the utterance's likelihood based on language models [Bangalore and Rambow, 2000, Langkilde-Geary, 2002], as well as sentence planning preferences from user ratings of output utterances [Walker et al., 2002, Stent et al., 2004]. While previous systems focus only on the most positively rated utterances (i.e. with the highest probability [Langkilde-Geary, 2002] or the highest estimated user rating [Walker et al., 2002]), we use the overgenerate and select method to hit any scalar target along the Big Five scales (1...7). Hence, the objective function used in PERSONAGE-OS is the inverse of the distance between the target personality scores and the scores predicted by the models.

Isard et al. [2006] extend Langkilde-Geary's method to control the generator's output personality by ranking utterances using n-gram language models trained on personality-annotated weblog data. Their CRAG-2 generator models personality by discretising the ratings into three groups (low, medium and high) resulting in three distinct models for each trait. Each model estimates the likelihood of the utterance given the personality type. In order to allow for more fine-grained control of the output variation, this chapter extends this work by building *continuous* models trained on user feedback that predict the personality scores of the candidate utterances, for all Big Five traits. Additionally, while CRAG-2 has yet to be evaluated, this chapter presents an evaluation of the learnt models on unseen utterances.

In the first part of this chapter, we focus on models trained on personality ratings of the random utterances presented in the previous chapter, i.e. two judges filled the Ten-Item Personality Inventory [Gosling et al., 2003] for a total of 160 utterances generated with uniformly random parameter values, from 20 content plans (for extraversion, three judges rated 320 utterances from 40 content plans). A subset of these utterances can be found in Appendix B. Our approach—illustrated in Figure 8.1—can be summarised as follows:

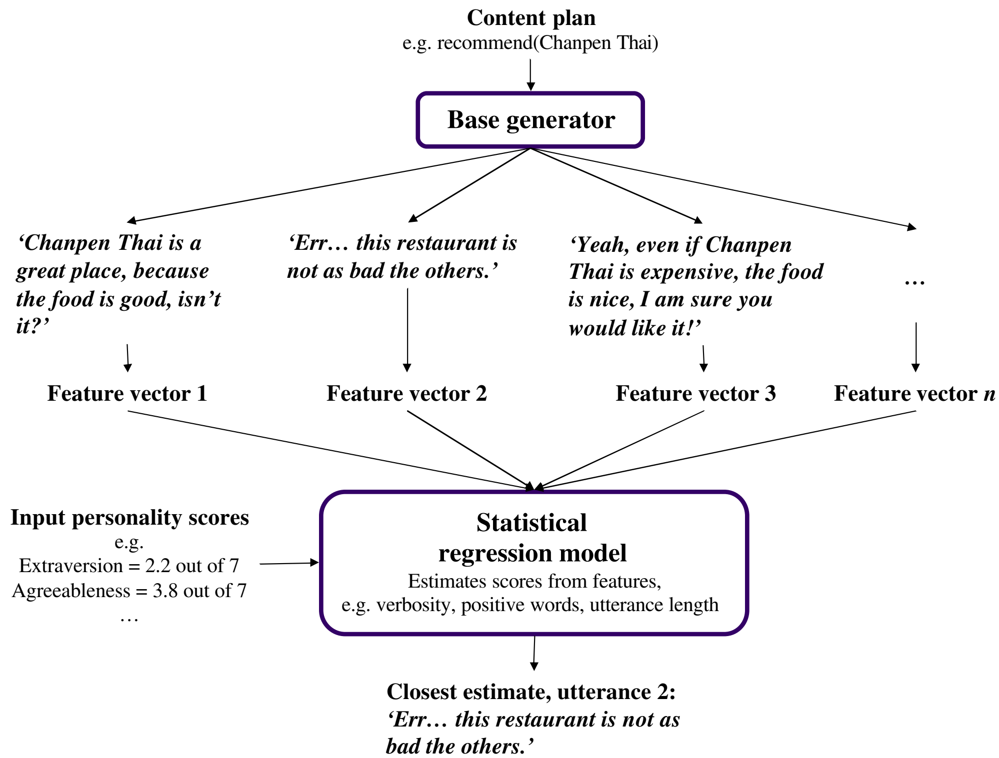


Figure 8.1: Illustration of PERSONAGE-OS's overgenerate and select method at generation time.

- At development time:

  1. Generate utterances covering the full parameter range (see Chapter 7);
  2. Judges rate the output with a standard personality test;
  3. Compute feature values for each utterance (see Chapter 7);
  4. Train a statistical model to predict the judges' ratings from the features of unseen utterances.

- At generation time:

  1. Generate candidate utterances covering the full parameter range; the number of utterances is dependent on processing capabilities and real-time dialogue constraints;
  2. Estimate the Big Five personality scores of each utterance using the statistical models trained at development time;

3. Output the utterance with the closest estimates to the target personality scores, e.g. according to the Euclidean distance.

While the main disadvantage of this approach is the computational cost of the overgeneration phase, it allows for the control of features that are emergent properties of multiple generation decisions, such as the utterance's length or instantiations of word categories. In Section 7.2, we discussed in full the types of features that can be used to characterise candidate utterances. We will make use of those features here. Section 8.2 describes the learning algorithms used to train our statistical models, and Section 8.3 evaluates these learning algorithms with each feature set.

8.2 Statistical models

Previous work on the overgenerate and select paradigm focuses on selecting the utterance maximising a constant objective function, such as the utterance's likelihood according to a language model [Langkilde-Geary, 2002] or an estimate of the utterance's quality based on user ratings [Walker et al., 2002]. These objective functions are best estimated by a *ranking* model, as only one or a few of the most highly rated utterances are of interest. However, controllable personality variation relies on an objective function that is dependent on input target personality scores. Absolute personality scores therefore need to be estimated for each candidate utterance, in order to be compared to the target scores. For this reason, we use *regression* models for the selection phase.

We use the Weka toolbox [Witten and Frank, 2005] for training and evaluation. It is not clear what learning algorithm is more appropriate for modelling personality. We therefore compare five algorithms with a baseline model returning the mean personality score in the training data (Base). The baseline model returns a constant value, which is equivalent to selecting the candidate utterance at random. We compare together a linear regression model (LR), an M5' regression tree (M5R), an M5' model tree returning linear models (M5), and support vector machines with a linear kernel (SVM_1) and a radial-basis function kernel (SVM_r). These learning algorithms were already used in Chapter 3 for training personality recognition models. Parameters of the algorithms are set to Weka's default values.

8.3 Results with in-domain models

The overgeneration and select approach requires a model that can select a small number of specific utterances among a set of candidate utterances; such selection models can be learnt statistically from personality-annotated data. We first evaluate models trained on PERSONAGE's output utterances (in-domain), while the next section investigates whether out-of-domain models trained on more general corpora are useful.

In order to investigate possible sources of error, the overall error between the target personality and the actual personality of the output utterance is broken down into two components: the *modelling error* and the *sampling error*, such that

$$\text{Overall error} = \text{modelling error} + \text{sampling error}$$

Each error component is analysed separately in the following sections, using selection models trained on PERSONAGE's random output utterances.

8.3.1 Modelling error

The modelling error results from the inaccuracy of the statistical selection model, i.e. assuming that the candidate utterance set contains an utterance matching the target personality exactly, is the model likely to select it? Regression models are evaluated using the correlation and the mean absolute error between the predicted scores and the judges' average ratings. Tables 8.1 and 8.2 respectively show the correlation and the mean absolute error on a scale from 1 to 7 over ten 10-fold cross-validations, for each learning algorithm and feature set. Models are trained on each of the following feature sets (detailed in Section 7.2): generation decisions (Gen), LIWC features, MRC features, n-gram frequency counts (Ngram), all content-related features (Content, i.e. generation decisions, LIWC and MRC), as well as all features together (All). To reduce data sparsity, we only select features that correlate significantly with the ratings with a coefficient higher than 0.1 ($p < .10$).

Early experiments showed that n-gram features overfit significantly when selected on the full dataset before cross-validation, producing misleadingly high accuracies (e.g. the SVM_r model trained in such a way produced a correlation of .66,

Trait	Feature set	**Base**	**LR**	**M5R**	**M5**	**SVM$_1$**	**SVM$_r$**
Extraversion	Gen	0.00	0.37 •	0.27 •	0.37 •	0.34 •	0.40 •
	LIWC	0.00	**0.45** •	0.21 •	0.45 •	0.43 •	0.35 •
	MRC	0.00	-0.03	-0.01	-0.04	0.00	0.00
	Ngram	0.00	0.06 •	0.13 •	0.38 •	0.09 •	0.23 •
	Content	0.00	0.38 •	0.24 •	0.38 •	0.37 •	0.40 •
	All	0.00	0.09 •	0.18 •	0.37 •	0.17 •	0.33 •
Emotional stability	Gen	0.00	0.19 •	0.18 •	0.25 •	0.22 •	**0.32** •
	LIWC	0.00	0.27 •	0.23 •	0.27 •	0.28 •	0.21 •
	MRC	0.00	-0.04	0.00	-0.04	-0.04	-0.04
	Ngram	0.00	-0.01	-0.01	0.10 •	0.03	0.07 •
	Content	0.00	0.25 •	0.13 •	0.30 •	0.23 •	0.31 •
	All	0.00	0.06 •	0.12 •	0.27 •	0.11 •	0.29 •
Agreeableness	Gen	0.00	0.29 •	**0.54** •	0.43 •	0.34 •	0.49 •
	LIWC	0.00	0.35 •	0.26 •	0.36 •	0.35 •	0.28 •
	MRC	0.00	0.11 •	0.29 •	0.21 •	0.13 •	0.14 •
	Ngram	0.00	0.03	0.07 •	0.28 •	0.23 •	0.35 •
	Content	0.00	0.24 •	0.52 •	0.43 •	0.30 •	0.49 •
	All	0.00	0.18 •	0.52 •	0.45 •	0.20 •	0.43 •
Conscientiousness	Gen	0.00	0.08 •	0.15 •	0.15 •	0.06 •	0.09 •
	LIWC	0.00	0.24 •	0.21 •	0.24 •	0.16 •	0.20 •
	MRC	0.00	0.07 •	0.00	0.08 •	0.08 •	0.08 •
	Ngram	0.00	-0.02	-0.03	0.13 •	-0.02	0.05
	Content	0.00	0.13 •	0.14 •	**0.25** •	0.09 •	0.11 •
	All	0.00	-0.03	0.12 •	0.19 •	-0.02	0.08 •
Openness to experience	Gen	0.00	0.00	-0.01	0.08 •	0.01	0.07 •
	LIWC	0.00	**0.19** •	0.09 •	0.16 •	0.15 •	0.16 •
	MRC	0.00	0.14 •	-0.01	0.17 •	0.16 •	0.18 •
	Ngram	0.00	0.05	0.00	0.10 •	0.04	0.12 •
	Content	0.00	0.07 •	0.05 •	0.17 •	0.08 •	0.13 •
	All	0.00	0.05	0.04 •	0.16 •	0.05 •	0.07 •

• statistically significant improvement
over the mean value baseline ($p < .05$)

Table 8.1: Pearson's correlation coefficients between the ratings and the predictions of selection models trained on different feature sets (Content = Gen + LIWC + MRC). Models are detailed in Section 8.2. All results are averaged over ten 10-fold cross-validations, and for each run features that do not correlate above .1 with the ratings in the training folds are removed. Best results for each trait are in bold.

compared to .23 in Table 8.1 with n-grams computed on the training data). Thus, the results presented here are with features repetitively filtered based on their correlation with ratings within the cross-validation *training* folds, i.e. not including the test data.

Table 8.1 shows that—in PERSONAGE's domain—extraversion is best modelled with LIWC features, with a correlation of .45 using either a linear regression model or a regression tree. The model trained on the full dataset is shown in Figure 8.2. Paired t-tests over the cross-validation folds show that these results improve significantly over the mean value baseline ($p < .05$, two-tailed). Table 8.2 shows that both the linear model and the regression tree yield a mean absolute error of .89 on

Trait	Feature set	**Base**	**LR**	**M5R**	**M5**	**SVM$_1$**	**SVM$_r$**
Extraversion	Gen	1.00	0.97 •	0.96 •	0.94 •	0.99	0.89 •
	LIWC	1.00	**0.89** •	0.97 •	**0.89** •	0.91 •	0.93 •
	MRC	1.00	1.01	1.00	1.02	1.02	1.01
	Ngram	1.00	1.73	0.99	0.94 •	1.21	0.97 •
	Content	1.00	0.97 •	0.96 •	0.93 •	0.98	0.90 •
	All	1.00	2.57	0.97 •	0.95 •	1.25	0.93 •
Emotional stability	Gen	0.99	1.12	0.96 •	0.98	1.11	**0.94** •
	LIWC	0.99	**0.94** •	**0.94** •	0.95 •	0.95 •	0.96 •
	MRC	0.99	0.99	0.99	0.99	1.06	1.04
	Ngram	0.99	1.50	0.99	1.06	1.25	1.01
	Content	0.99	1.12	0.97 •	0.97	1.14	0.95 •
	All	0.99	1.90	0.97	0.99	1.38	0.94 •
Agreeableness	Gen	0.86	0.97	**0.70** •	0.76 •	0.91	0.72 •
	LIWC	0.86	0.80 •	0.82 •	0.80 •	0.81 •	0.82 •
	MRC	0.86	0.88	0.81 •	0.86	0.89	0.85 •
	Ngram	0.86	2.34	0.86	0.84	0.98	0.77 •
	Content	0.86	1.12	0.71 •	0.78 •	1.00	0.71 •
	All	0.86	1.01	0.71 •	0.77 •	1.01	0.74 •
Conscientiousness	Gen	0.89	1.14	0.88 •	0.94	1.21	0.97
	LIWC	0.89	0.89	0.89	0.89	0.93	**0.87** •
	MRC	0.89	0.90	0.89	0.90	0.91	0.91
	Ngram	0.89	1.58	0.90	0.95	1.17	0.95
	Content	0.89	1.17	0.90	0.93	1.21	0.96
	All	0.89	1.55	0.90	0.95	1.39	0.98
Openness to experience	Gen	1.00	1.17	1.00	1.06	1.23	1.03
	LIWC	1.00	0.99	0.99	1.00	1.02	0.98 •
	MRC	1.00	0.98 •	1.00	**0.97** •	0.99	0.99 •
	Ngram	1.00	1.53	1.00	1.06	1.18	0.99
	Content	1.00	1.21	1.00	1.03	1.21	1.01
	All	1.00	1.52	1.00	1.03	1.26	1.04

• statistically significant improvement
over the mean value baseline ($p < .05$)

Table 8.2: Mean absolute error between the ratings and the predictions of selection models trained on different feature sets (Content = Gen + LIWC + MRC). Models are detailed in Section 8.2. All results are averaged over ten 10-fold cross-validations, and for each run features that do not correlate above .1 with the ratings in the training folds are removed. Best results for each trait are in bold.

a scale from 1 to 7, whereas the baseline makes an average error of 1.00. Generation decision features also perform well, with a correlation of .40 using support vector machines. Interestingly, increasing the feature set by combining generation decisions with LIWC features does not improve accuracy, suggesting that the level of redundancy between the feature sets outweighs their complementarity. While MRC features do not outperform the baseline, n-gram features produce a significant correlation of .38 using a model tree, but they do not perform as well as LIWC categories.

Emotional stability is more difficult to model than extraversion, with a maximum correlation of .32 using a support vector machine model trained on genera-

```
Extraversion =

     0.0288 * LIWC - Word count +
    -0.0063 * LIWC - Words per sentence +
    -0.0187 * LIWC - Period +
     0.3389 * LIWC - Exclamation marks +
    -0.0204 * LIWC - All punctuation +
    -0.3396 * LIWC - Negations +
    -0.0705 * LIWC - Assents +
     0.0969 * LIWC - Optimism +
    -0.0021 * LIWC - Negative emotion +
    -0.0800 * LIWC - Tentative +
    -0.1154 * LIWC - Time +
     0.0050 * LIWC - Achievement +
     0.0320 * LIWC - Physical states +
     0.0035 * LIWC - Dictionary +
     4.4340
```

Figure 8.2: Linear regression model predicting extraversion from PERSONAGE's utterances, trained on LIWC features. See Table 3.5 in Chapter 3 for details on LIWC features.

tion decision features, which corresponds to a mean absolute error of .94 compared with a .99 baseline error. Again, generation decision and LIWC features perform best, whereas MRC features never outperform the baseline. Correlations of n-gram models never exceed .10, confirming the hypothesis that neuroticism is manifested by content-level features rather than surface markers [Oberlander and Gill, 2004b].

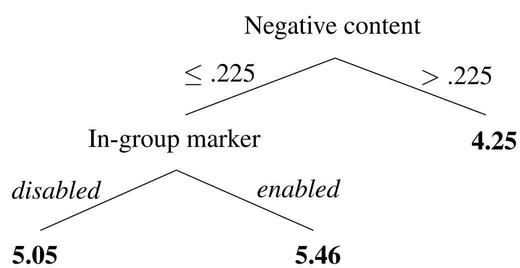


Figure 8.3: M5' regression tree predicting agreeableness from PERSONAGE's utterances, trained on generation decision features.

Agreeableness models produce the highest accuracy among all traits, with a .54 correlation using a regression tree with generation decision features, i.e. a .70 mean error compared with a .86 baseline error (19% decrease). Figure 8.3 shows that the learnt model is relatively simple, with only two features and three possible outputs. Generation decision features perform best across models, followed by

LIWC features, however MRC features also perform well with correlations up to .29. N-gram frequency counts also outperform the baseline with correlations up to .35. Although combining all features together does not improve performance, these results suggest that agreeableness can be modelled successfully at different levels of utterance representation.

Whereas agreeableness models produce encouraging results, conscientiousness is more difficult to model, with a maximum correlation of .25 using a model tree with all content features together. The model tree trained on the full dataset is pruned down to one node, thus it is actually equivalent to the linear regression model shown in Figure 8.4. LIWC features produce the best results on their own, followed by generation decisions. Although models based on MRC or n-gram features both outperform the baseline, they correlate more weakly with the judges' ratings (correlation below .14).

```
Conscientiousness =

    -0.75790 * GEN - Content planning - repetitions
    - 1.2099 * GEN - Content planning - negative content
    - 0.4580 * GEN - Aggregation - justify - so cue word
    + 0.6526 * GEN - Aggregation - infer - merge
    - 1.1956 * GEN - Aggregation - concede - although cue word
    + 2.0226 * GEN - Lexical choice - lexicon frequency
    + 2.5381 * GEN - Lexical choice - lexicon word length
    - 1.4956 * GEN - Lexical choice - maximum lexicon frequency
    - 0.5802 * GEN - Lexical choice - minimum lexicon word length
    + 0.5485 * GEN - Request confirmation: let's see what
    - 0.0079 * LIWC - Question marks
    - 0.1362 * LIWC - Negative emotion
    + 0.0338 * LIWC - Social processes
    + 0.0033 * MRC - Concreteness
    + 4.0868
```

Figure 8.4: Linear regression model predicting conscientiousness from PERSONAGE's utterances, trained on generation decision, LIWC and MRC features (degenerated model tree).

Finally, openness to experience is the hardest trait to model in our domain, with a maximum correlation of .19 using the linear regression model with LIWC features in Figure 8.5. Interestingly, MRC features perform comparatively well with a correlation of .18 using support vector machines. Additionally, Table 8.2 shows that the model tree trained on MRC features produces the lowest mean absolute error (.97). Although they improve on the baseline, generation decision features only produce a correlation of 0.10 with the judges' ratings.

```
Openness to experience =

     -0.0964 * LIWC - Question marks +
      0.0664 * LIWC - Exclamation marks +
      0.2970 * LIWC - We +
     -0.0664 * LIWC - Assents +
      0.0246 * LIWC - Prepositions +
      0.0717 * LIWC - Numbers +
     -0.0268 * LIWC - Affective processes +
     -0.0424 * LIWC - Positive feeling +
      0.1583 * LIWC - Optimism +
     -0.0429 * LIWC - Negative emotion +
      0.0384 * LIWC - Social processes +
      0.0639 * LIWC - Friends +
      0.0043 * LIWC - Up +
      0.1610 * LIWC - Sexuality +
      3.6763
```

Figure 8.5: Linear regression model predicting openness to experience from PERSONAGE's utterances, trained on LIWC features.

8.3.1.1 Discussion

The results presented here show that models of each Big Five trait successfully outperform the baseline in PERSONAGE's information presentation domain. Perceptions of agreeableness and extraversion are easier to model, whereas conscientiousness and openness to experience are more difficult. A possible explanation is that these traits are not conveyed well in PERSONAGE's narrow domain. However, personality recognition results in Chapter 3 suggest that observed openness to experience is difficult to model using general conversational data as well. It is therefore possible that this trait may not be expressed through spoken language as clearly as other traits.

Over all traits, Table 8.1 shows that generation decision and LIWC features generalise better to the test data than n-gram frequency counts. This suggests than even in a narrow domain, the specificity of n-gram features is detrimental to performance on unseen utterances. Even when added to other features, they decrease the overall accuracy as the learning algorithm is more likely to pick features that do not generalise to the test data. It is possible that individual n-grams improve performance, but their identification requires evaluating the models with each feature added separately, which we leave as future work. The usefulness of MRC features is highly dependent on the personality trait, as they predict openness to experience successfully, but not extraversion and emotional stability.

Concerning the learning algorithms, results show that personality can generally be modelled with simple models, as linear regression models perform best for extraversion, conscientiousness and openness to experience. Additionally, the simple regression tree in Figure 8.3 outperforms all other learning algorithms for predicting agreeableness. Although the SVM model produces the best results for predicting emotional stability, linear regression models also correlate significantly with the judges' ratings of that trait (with a 21% performance decrease).

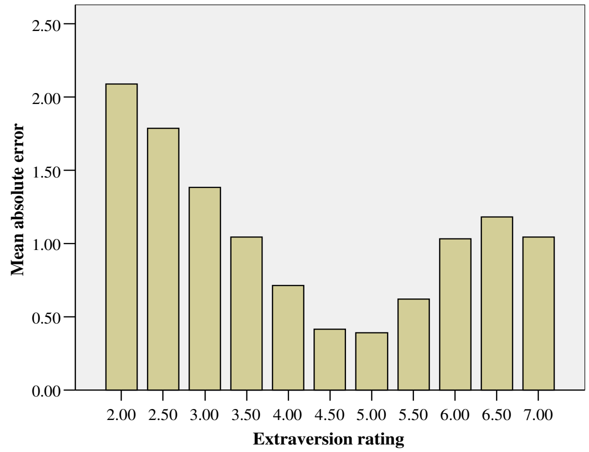
The best models produce correlation coefficients ranging from .19 (openness to experience) to .54 (agreeableness). These correlations are comparable to the level of agreement between pairs of judges on random utterances (see Section 7.1.2), suggesting that the selection models predict the average rating as well as a new judge.

8.3.1.2 Modelling error distribution

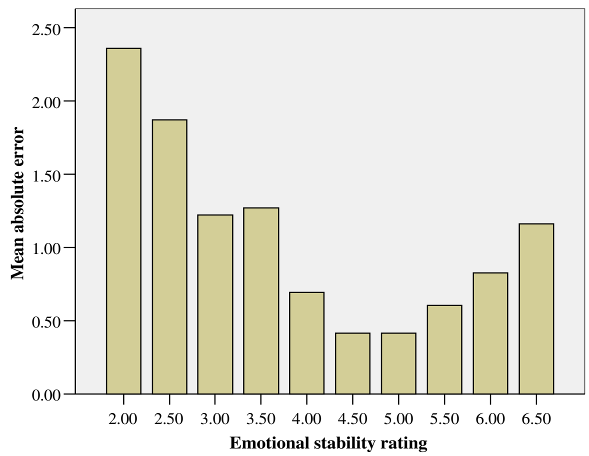
Although PERSONAGE-OS's personality models correlate significantly with the judges' ratings, it is not clear whether models perform well for any target personality score. Thus, this section analyses how prediction accuracy varies for different ratings.

Figures 8.6 and 8.7 show the distribution of the mean absolute error from a 10-fold cross-validation of the models yielding the highest overall correlation with the ratings. A consistent result over all traits is that extreme utterances are predicted less accurately than utterances targeted to the mid-range of the scale (4 out of 7). This bias can be interpreted as a data sparsity problem, as there is only a few number of extreme utterances to learn from due to the normal distribution of the random utterance ratings (see Section 7.1). Also, the large number of utterances rated as neutral suggests that results in Table 8.1 are likely to favour models that perform well on those non-extreme utterances.

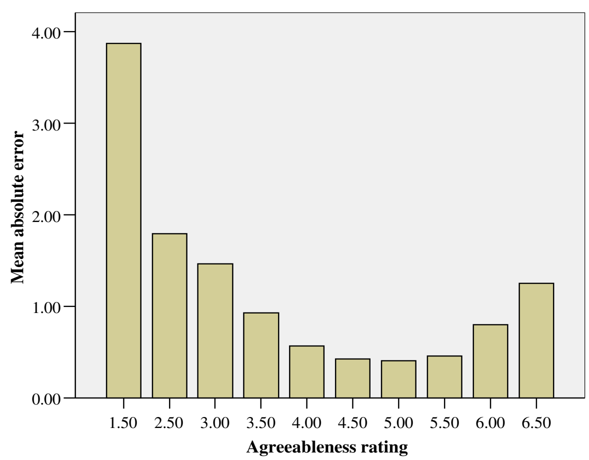
A second observation is that the selection models are usually better at predicting personality at one end of the scale: extraversion, emotional stability, agreeableness and conscientiousness are modelled more successfully than introversion, neuroti-



(a) Extraversion (linear regression with LIWC features)

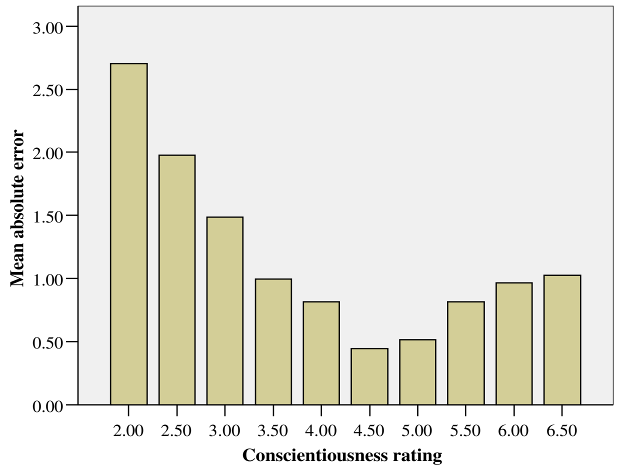


(b) Emotional stability (SVM with generation features)

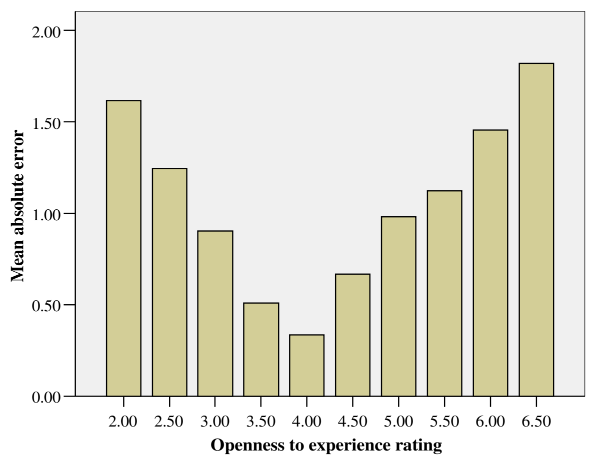


(c) Agreeableness (regression tree with generation features)

Figure 8.6: Error distribution of the best performing models, averaged over bins of width .5. Individual predictions are obtained over a 10-fold cross-validation.



(a) Conscientiousness (model tree with content features)



(b) Openness to experience (linear regression with LIWC features)

Figure 8.7: Error distribution of the best performing models, averaged over bins of width .5. Individual predictions are obtained over a 10-fold cross-validation.

cism, disagreeableness and unconscientiousness. For example, the distribution for extraversion in Figure 8.6(a) shows that on average the model produces an error above 1.75 out of 7 when targeting extraversion scores below 2.75. The positive end of the extraversion scale is modelled more accurately, although the model still makes an average error of around 1 out of 7 when targeting scores above 6.

Results in this section suggest that models do not learn extreme utterances as well as neutral ones. A possible solution is to make the error distribution more uniform by giving more importance to extreme utterances during the learning process, which is typically done by increasing the weight of extreme instances in the dataset.

The weight distribution could assign a weight that is inversely proportional to the ratings' density, or proportional to the modelling error reported in Figures 8.6 and 8.7. We leave this optimisation as future work, as it is not clear (1) what the optimal weighting function should be, and (2) what its effect would be on the overall accuracy.

8.3.2 Sampling error

The overgeneration phase can be interpreted as a sampling problem in the variation space, therefore a sampling error is observed if the candidate utterance set does not contain an utterance predicted to have the target personality. This error is the distance between the target personality score and the predicted score of the closest utterance in the candidate utterance set, which is independent of the model's accuracy at predicting the judges' ratings (modelling error). Given perfect selection models, the sampling error would be the only error component of the overall generator. Figure 8.8 illustrates the sampling error for a candidate utterance set containing two utterances. Although the true rating of utterance A is 3 out of 7, its rating is estimated to be 5 by the model (modelling error = 2.0), which is the closest to the target score (6.5). Thus, utterance A is selected over utterance B and the resulting sampling error is 1.5 out of 7 (total error = 3.5).

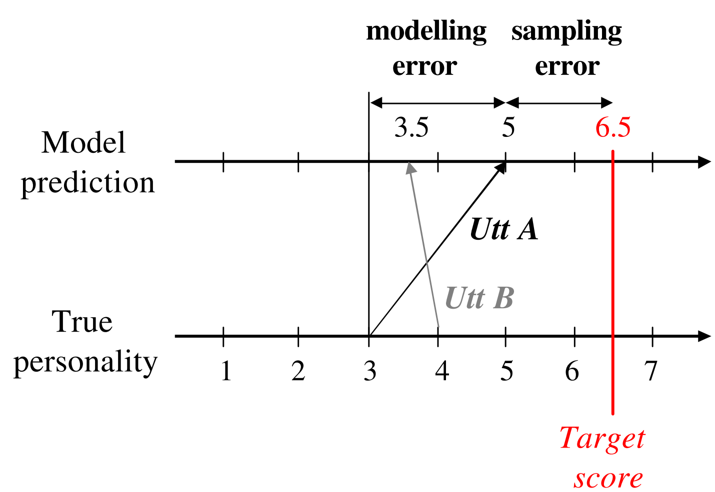


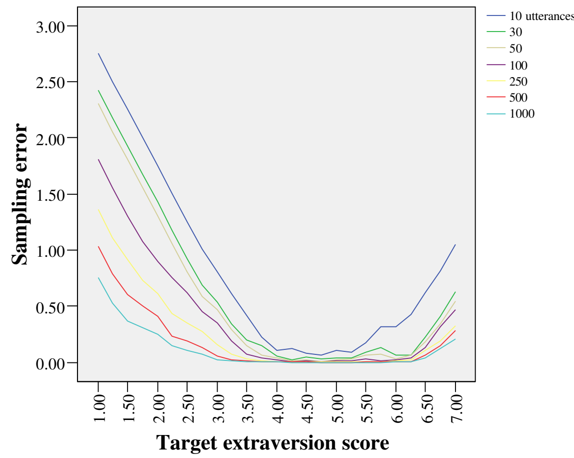
Figure 8.8: Modelling error and sampling error for a candidate utterance set of size 2. Utterance A is selected because its predicted score is closer to the target, yielding a sampling error of 1.5.

As most previous work on statistical language generation only focuses on one end of the selection scale, this error component is typically ignored [Langkilde-Geary, 2002, Walker et al., 2002], with the exception of Paiva's work [2004]. The sampling error can be reduced by (1) increasing the number of candidate utterances, and (2) ensuring that the predicted ratings of the candidate utterances are uniformly—or at least widely—distributed.

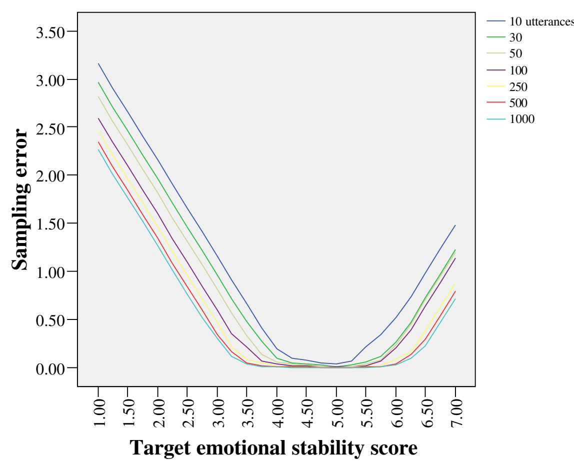
Regardless of the true personality of the selected utterance, the sampling error can be estimated by generating candidate utterances expressing a given content plan, and computing the distribution of the distance between predicted score of the model's selection over all possible target scores. Figures 8.9 and 8.10 show this distribution for different numbers of candidate utterances. An objective of this study is to assess the optimal number of candidate utterances, which requires a trade-off between sampling error and overgeneration time. As some content plans produce utterances that are easier to model than others, results are averaged over 10 randomly selected content plans. Results were obtained using the best performing models according to the cross-validation results in Table 8.1 in the previous section.

Figure 8.9(a) shows that with a linear regression model, increasing the number of candidate utterances reduces the distance between the model prediction and the target score, over the full extraversion scale. Interestingly, introversion benefits the most from this increase. The largest error occurs when targeting a score of 1 out of 7, with an average error of 2.75 with 10 candidate utterances, which is reduced to .76 with 1000 utterances. Concerning the positive end of the extraversion scale, 100 utterances are enough for selecting an utterance closer than .5 to the target, for any target score above 2.75 out of 7. Utterances close to the mid-range produce almost no sampling error, as 30 utterances are enough for targeting scores between 3.25 and 6.25 with an average error below 0.5.

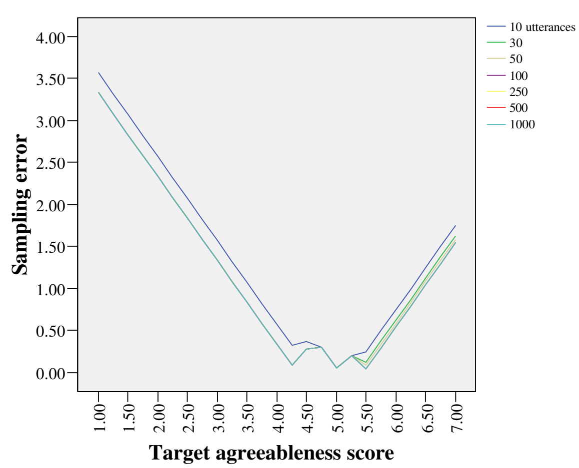
A similar pattern is observed in Figure 8.9(b) for emotional stability, with the SVM model trained on generation decision features. However, increasing the candidate utterance set is less beneficial than with linear regression, as doubling the utterance set size from 500 to 1000 only produces a marginal improvement. The sampling error for the middle and the positive end of the scale is greatly reduced with 250 utterances, as with this candidate set size the average error of the se-



(a) Extraversion (linear regression with LIWC features)

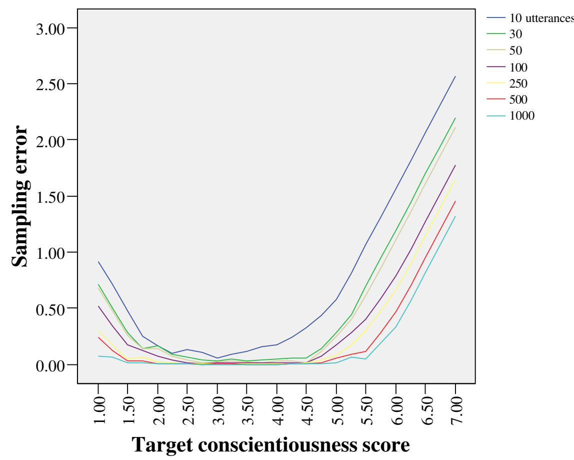


(b) Emotional stability (SVM with generation features)

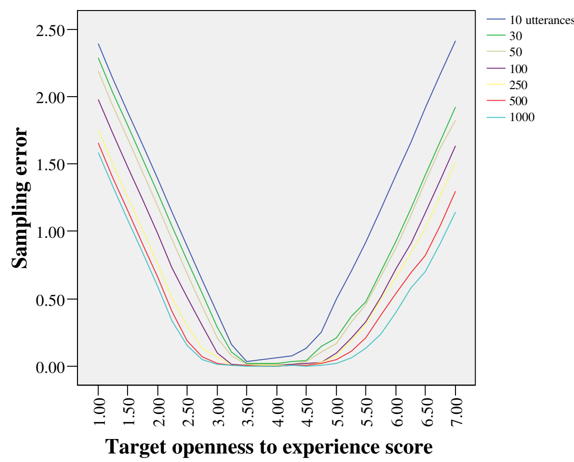


(c) Agreeableness (regression tree with generation features)

Figure 8.9: Distance of the selected utterance's predicted score to the target personality score (sampling error), for different numbers of candidate utterances and targets. Results are averaged over 10 input content plans.



(a) Conscientiousness (model tree with content features)



(b) Openness to experience (linear regression with LIWC features)

Figure 8.10: Distance of the selected utterance's predicted score to the target personality score (sampling error), for different numbers of candidate utterances and targets. Results are averaged over 10 input content plans.

lected utterance falls below 1.0 for any target emotional stability above 2.5 out of 7. Extreme neuroticism is harder to sample, as even 1000 utterances produce a large error (above 2.0) when targeting a score of 1 out of 7.

Although the agreeableness model produces the highest correlation with the judges' ratings (see Table 8.1), Figure 8.9(c) shows that the model does not make extreme predictions: 1000 utterances only ensure an average error below 1.0 for a target agreeableness between 3.50 and 6.25. Interestingly, the error is almost independent of the number of candidate utterances. This independence reflects the simplicity of the regression tree model illustrated in Figure 8.3, which only has three possible output values. Figure 8.9(c) illustrates the fact that once the candidate utterance set contains all possible model predictions (e.g. with 100 candidate utterances), there is no benefit in generating additional utterances. Furthermore, the linear slope reflects the lack of variation of the model outside the $[4.25, 5.46]$ interval. This confirms that models that provide a continuous variation of their output (e.g. linear regression) are better suited to the overgenerate and select task than models with a few discrete outputs, as they have more discriminative power, and are thus more likely to benefit from a larger candidate utterance set.

The linear regression tree predicting conscientiousness from content features does not suffer from that problem. It successfully covers the middle and the low end of the scale, as 500 utterances produce an average sampling error below 0.5 for any target score below 6 out of 7, while 1000 utterances reduce the error to a maximum of .33 over the same target range. The model does not assign extremely conscientious scores, with an average distance of 1.32 when targeting a conscientiousness score of 7 with 1000 utterances.

The linear regression model predicting openness to experience from LIWC features in Figure 8.10(b) covers the middle range of the openness to experience scale, with an average distance below 1.0 for target scores between 2 and 6 (inclusive) with 250 utterances. However, the model does not identify both extremes—even with 1000 utterances—as it produces a sampling error of 1.59 when targeting an openness to experience score of 1 out of 7, and 1.14 when targeting a score of 7. More generally, the benefits of a larger candidate utterance set is less clear than with previous linear regression models, as doubling the set size from 500 to 1000 only decreases the average distance by 0.15 or less over the full scale.

This section highlights an important factor for choosing the optimal selection model, which has been overlooked in previous work on the overgenerate and select paradigm. While Occam's razor suggests that—other things being equal—simpler selection models should be preferred, there is support for choosing models based on the *variation of the output space*, rather than solely on their predictive power. More precisely, models that hit a large, uniform range should be favoured. The regression tree predicting agreeableness in Figure 8.3 illustrates this requirement, as Figure 8.9(c) shows that the lack of variation in the model's output makes it inappropriate for discriminating between candidate utterances, even though it produces the lowest overall error on our dataset.

8.3.3 Psychologically informed selection models

Although Section 8.3.1 shows that personality models outperform the baseline (i.e. lower modelling error), *extreme* personality is harder to model because of the lack of random utterances with extreme ratings. See ratings distributions in Figures 7.1 and 7.2. In Chapter 6, we successfully modelled extreme personality using a rule-based approach with parameter settings derived from psycholinguistic studies. These results suggest that both methods are complementary, as—by definition—manifestations of personality at the two extremes of the scale are rare, since the scales are assumed to be normally distributed in the population. In this section, we investigate techniques that use knowledge from the rule-based approach to improve the generation of personality over the full rating scale.

A first approach would consist of a generator with two modes: (1) with predefined parameter settings and (2) overgenerate and select, that switches from the latter to the former when target scores are close to extreme values (1 or 7) and/or if the selection model's predicted error is too high. However, it is not clear how to map the rule-based utterances onto the continuous personality scale, as some random candidate utterances might be perceived as more extreme.

In order to address this issue, we propose an approach that takes into account both methods, by training a selection model for both randomly generated utterances and rule-based utterances conveying extreme traits. As the feature distribution differs considerably between the two utterance sets, a first method for merging datasets is to concatenate them and add a feature indicating how each utterance was generated, i.e. with a parameter setting expressing either the low end or the high end of a trait, or with random parameters. The statistical model can thus learn for what target scores it should favour rule-based utterances over random ones, e.g. by biasing utterances obtained with the introvert parameter setting towards introversion even if some of their features do not correlate with introversion according to the random data.

A second method considers the low, high and random parameter settings as three different domains. We can therefore apply domain-adaptation techniques to improve on the dataset concatenation approach. We use a state-of-the-art domain adaptation method presented by Daume [2007]. This technique merges both datasets by duplicating each feature for each domain (i.e. three times here), repetitively setting the duplicate feature value to zero for all but one domain. The underlying hypothesis is that this explicit duplication makes it easier for the learning algorithm to determine which features are useful in a specific domain.

Metric	**Correlation coefficient**			**Mean absolute error**		
Dataset	**Random**	**Concat**	**Dupl**	**Random**	**Concat**	**Dupl**
Extraversion	.45	**.64** •	.64 •	.89	.82 •	.82 •
Emotional stability	.32	**.66** •	.64 •	.94	.74 •	.75 •
Agreeableness	**.54**	.54	.52	.70	.76	.77
Conscientiousness	.25	**.39** •	.34 •	.93	.86 •	.90
Openness to experience	.19	**.31** •	.25 •	.99	.96	1.02

• statistically significant improvement over the
model trained on the random utterances (Random)

Table 8.3: Accuracies for models trained on both rule-based (10% low and 10% high) and random utterances (80%), using the concatenation with domain indicator feature method (*Concat*) and Daume's feature duplication technique (*Dupl*). Results are obtained over ten 10-fold cross-validations of the best performing models, and compared with the models trained on the random utterances in Section 8.3.1 (*Random*).

Results in Table 8.3 show that models trained on the concatenated dataset perform better than when trained on the random utterances only ($p < .05$).[1] For example, the correlation of the best model predicting extraversion with the judges' ratings increases from .45 to .64 when adding 80 rule-based utterances (20%) conveying both ends of the scale, together with a domain indicator feature. Daume's domain adaptation technique provides equivalent performance improvement. This result suggests that utterances generated from parameter settings derived from psychological studies are easier to model, and they also extend the coverage of the candidate utterance set. Significant improvements are also observed for emotional stability, conscientiousness and openness to experience, however not for agreeableness. Results suggest that the learning algorithms use the domain indicator feature successfully, whereas Daume's feature duplication method increases the size of the feature space with no clear benefits.

In this section, we have evaluated how the rule-based approach presented in Chapter 6 can enhance the coverage of PERSONAGE-OS for continuous personality

[1] Although a paired t-test was used, reported significance levels are only approximations as different cross-validation folds were used due to the different dataset sizes.

generation, by using a domain indicator feature which lets the model learn when rule-based utterances are preferable over random ones. A reason for this improvement is that data sparsity makes it difficult to learn to predict extreme personality scores from random utterances, whereas the addition of rule-based utterances considerably flattens the error distribution. Daume's more complex domain adaptation techniques does not perform better than the single domain indicator feature, which could result from the high similarity between domains.

8.4 Results with out-of-domain models

The previous section shows that it is possible to produce utterances expressing personality varying on a continuous scale, by learning models from human ratings of utterances in the generation domain. Although models perform significantly better than the baseline for all traits, it is not clear whether they would perform well in a new domain: even if one's personality does not vary across situations, the way personality affects behaviour does. If domain-specific models do not generalise, adapting to a new domain would require the collection of new training data, which is a time-consuming process. A solution is to train personality recognition models on general data such as detailed in Chapter 3, and re-use them across domains [Mairesse et al., 2007]. While we do not expect out-of-domain models to perform as well as domain-specific ones, this section evaluates (1) whether they are accurate enough to be re-used in a specific domain without additional training, and (2) whether out-of-domain data can be used together with in-domain data to improve overall performance.

The first evaluation assesses the accuracy of Chapter 3's domain-independent personality recognition models at predicting the personality of PERSONAGE's output utterances. These models were trained on daily-life conversational data using observer reports of personality, as well as on stream-of-consciousness essays using self-reports [Mairesse et al., 2007, Mehl et al., 2006, Pennebaker and King, 1999]. The evaluation of these models on unseen data from their own domain is presented in Chapter 3.[2] Out-of-domain models are trained on standardised data to improve

[2]Chapter 3 did not evaluate n-gram features because of the risk of overfitting. However, since other work has used n-grams in selection models [Isard et al., 2006], we include n-gram features that correlate with the ratings with a coefficient of .1 or more.

generalisation. Feature values of PERSONAGE's utterances are thus standardised over the full set of output utterances, i.e. with zero mean and unit standard deviation.

8.4.1 Out-of-domain model accuracy

			Observers/conversations			Self-reports/essays		
Trait	**Feature set**	**Base**	**LR**	**M5**	**SVM**	**LR**	**M5**	**SVM**
Extraversion	LIWC	.00	.20 •	**.25** •	.16 •	.05	.05	.12 •
	MRC	.00	-.08	-.08	-.12	.11	.11	.12 •
	Ngram	.00	.16 •	.08	.14 •	.05	.03	.05
	All	.00	.20 •	.16 •	.20 •	.14 •	.13 •	.11 •
Emotional stability	LIWC	.00	.04	-.05	-.10	-.13	-.16	-.19
	MRC	.00	-.10	-.10	-.10	-.06	-.06	-.06
	Ngram	.00	-.01	**.05**	-.04	.01	.00	-.02
	All	.00	.02	-.05	.00	-.02	.03	-.05
Agreeableness	LIWC	.00	.00	-.03	-.10	.30 •	.30 •	**.33** •
	MRC	.00	-.16	-.11	-.28	.04	.04	.05
	Ngram	.00	-.19	-.22	-.15	.01	.05	.05
	All	.00	-.19	.02	-.13	.17 •	.18 •	.16 •
Conscientiousness	LIWC	.00	.12	.13	.10	.19 •	.19 •	**.20** •
	MRC	.00	-.11	-.11	-.13	-.05	-.05	-.02
	Ngram	.00	-.07	-.15	-.12	.08	.08	-.02
	All	.00	.03	.11	-.01	.11	.18 •	-.03
Openness to experience	LIWC	.00	-.11	-.08	-.05	-.14	-.14	-.09
	MRC	.00	.09	.09	.05	-.09	-.09	-.08
	Ngram	.00	-.02	.00	-.02	.08	**.11**	.10
	All	.00	-.05	-.06	-.05	.07	.01	.10

• statistically significant improvement
over the mean value baseline ($p < .05$, two-tailed)

Table 8.4: Correlation between out-of-domain model predictions and judges' ratings over the random utterances, for different feature sets and models.

Table 8.4 shows the correlation between out-of-domain model predictions and the judges' ratings over all random utterances, for different models and feature sets. Results show that the best model for extraversion is the M5' model tree with LIWC features ($r = .25$), confirming that LIWC features generalise better than n-grams for out-of-domain models as well as in-domain models. Surprisingly, models trained on self-reports perform best for agreeableness and conscientiousness, with correlations of .33 and .20 using an SVM model with LIWC features. This finding implies that self-reports of personality can be more accurate at predicting *observed* personality in a new domain. Whether this performance increase is due to the specific nature of stream-of-consciousness essays or to the larger amount of training data remains to be evaluated. As with in-domain models, agreeableness models

			Observers/conversations			Self-reports/essays		
Trait	**Feature set**	**Base**	**LR**	**M5**	**SVM**	**LR**	**M5**	**SVM**
Extraversion	LIWC	.00	.29 •	**.33** •	.21 •	.07	.07	.17 •
	MRC	.00	-.07	-.07	-.12	.02	.02	.06
	Ngram	.00	.18 •	.05	.19 •	.11 •	.04	.16 •
	All	.00	.27 •	.29 •	.28 •	.24 •	.19 •	.22 •
Emotional stability	LIWC	.00	**.16** •	.12 •	.05	-.46	-.46	-.50
	MRC	.00	.03	.03	-.03	-.16	-.16	-.20
	Ngram	.00	-.07	.15 •	-.11	-.27	-.35	-.33
	All	.00	-.04	.01	-.09	-.33	-.28	-.33
Agreeableness	LIWC	.00	.12 •	.14 •	.09	.31 •	.31 •	**.36** •
	MRC	.00	-.10	-.06	-.22	.10	.10	.11
	Ngram	.00	.01	-.08	.04	.06	.13 •	.06
	All	.00	.07	.22 •	.12 •	.25 •	.16 •	.17 •
Conscientiousness	LIWC	.00	.19 •	.19 •	.23 •	.17 •	.17 •	.21 •
	MRC	.00	.01	.01	.02	-.02	-.02	.04
	Ngram	.00	.04	-.05	-.03	.16 •	.15 •	.07
	All	.00	.10	.18 •	.04	.25 •	**.26** •	.11
Openness to experience	LIWC	.00	-.17	-.16	-.10	-.17	-.17	-.05
	MRC	.00	.12 •	.12 •	**.13** •	-.16	-.16	-.13
	Ngram	.00	.03	-.20	.01	.04	.11	.03
	All	.00	.03	.03	-.02	-.07	-.16	-.10

• statistically significant improvement
over the mean value baseline ($p < .05$, two-tailed)

Table 8.5: Correlation between the out-of-domain model predictions and judges' ratings over both random and rule-based utterances, for different feature sets and models. Within the domain adaptation task, this represents a baseline using only the source data.

produce the highest accuracy, whereas out-of-domain models for emotional stability and openness to experience do not significantly outperform the baseline.

In order to evaluate whether out-of-domain models can also predict the personality of utterances generated using the rule-based approach, Table 8.5 evaluates the effect of adding Chapter 6's rule-based utterances to the candidate utterance set. As with in-domain data, results show that the out-of-domain models perform better with the rule-based utterances for all traits. Table 8.6 summarises results with the best models for each trait, showing that although accuracies are worse than with in-domain data (see Section 8.3), out-of-domain models outperform the random baseline for all traits. LIWC features tend to generalise best to the new domain for extraversion, emotional stability and agreeableness, whereas n-gram features generalise poorly. As with in-domain models, MRC features are most useful for predicting observed openness to experience.

Although out-of-domain models perform worse than models trained on the output utterances, their accuracies are surprisingly high given (1) the important differ-

Trait	Model	Type	Features	r_{best}	e_{best}	e_{base}
Random utterances only:						
Extraversion	M5	Obs	LIWC	.25 •	.86	.80
Emotional stability	M5	Obs	Ngram	.05	.90	.84
Agreeableness	SVM	Self	LIWC	.33 •	.79 •	.83
Conscientiousness	SVM	Self	LIWC	.20 •	.81	.83
Openness to experience	M5	Self	Ngram	.11	.85	.86
Random and extreme utterances:						
Extraversion	M5	Obs	LIWC	.33 •	.81	.83
Emotional stability	LR	Obs	LIWC	.16 •	.94	.85
Agreeableness	SVM	Self	LIWC	.36 •	.77 •	.82
Conscientiousness	M5	Self	All	.26 •	.82 •	.85
Openness to experience	SVM	Obs	MRC	.13 •	.89	.86

• statistically significant improvement over the mean value baseline, with error e_{base} and null correlation ($p < .05$, two-tailed)

Table 8.6: Summary of the correlation (r_{best}) and absolute error (e_{best}) between the judges' personality ratings and the predictions of the best out-of-domain models, compared with the error e_{base} of the mean value baseline (computed on all in-domain data). Type *Obs* indicates a model trained on observer reports of conversations, while *Self* indicates a model trained on self-report and essays. Feature values and scores are standardised, as well as the absolute errors. Significance tests compare the error made on individual instances.

ences between domains; (2) the small training set size (96 instances for conversation models); (3) the fact that the judgements were made over a single utterance; and (4) the low inter-rater correlation on the random utterances. For extraversion, the correlation of the best out-of-domain model with the average ratings is only .05 lower than the inter-rater correlation shown in Table 7.3.

8.4.2 Domain adaptation

Out-of-domain models perform better than the baseline, but can they be combined with in-domain data to improve overall performance? In this section, we assess whether out-of-domain models of personality recognition—trained on the general *source* domain—can be re-used in various applications combined with a small amount of data from the *target* domain.

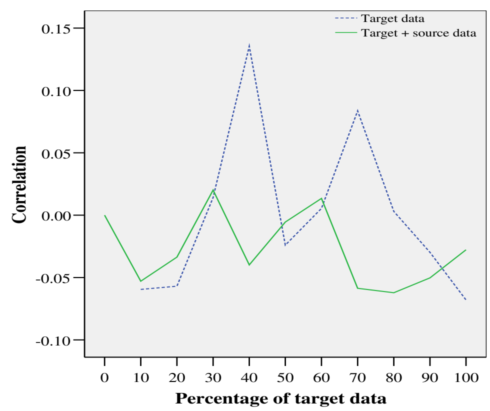
To answer this question, we apply a state-of-the-art domain adaptation technique presented by Daume [2007], previously used in Section 8.3.3 for combining random and rule-based utterances. This method merges both datasets by duplicating each feature twice, setting the first duplicate to zero for instances in the source domain, and setting the second one to zero for instances in the target domain. The underlying hypothesis is that this explicit duplication makes it easier for the learning algorithm to determine what feature is useful in the target domain.

For each trait, the source data consists of the dataset that yields the best performance according to Table 8.4, i.e. either the conversation extracts and observer reports or the essays and self-reports used in the previous section. The target data consists of randomly generated utterances and the judges' ratings. To evaluate the contribution of the source data, the target data is randomly partitioned into a training set as large as the source data, and a test set containing the remaining utterances. Features with a correlation coefficient below .1 with the personality scores are removed from both the source and target datasets. From the target data's training set, 10 subsets of various sizes are successively merged with the source data. A statistical model is then trained on each merged subset, as well as without the source data. Results are averaged over 10 random initial partitions. Figures 8.11 and 8.12 compare accuracies using the best performing out-of-domain models (according to Table 8.4) trained with and without source data, for various training set sizes.

Results in Figure 8.11(a) show that the model tree predicting extraversion (trained on LIWC features) benefits from domain adaptation, as the addition of source data improves accuracy for any amount of target data up to 90% of the source data. Results for other traits show that domain adaptation does not improve performance over models trained only on target data. While Figure 8.11(b) reports low accuracies for models of emotional stability, the baseline accuracy reported in Table 8.4 suggests that our out-of-domain models perform poorly for that trait. Concerning the other traits, the lack of improvement could be explained by the nature of the source domain (i.e. self-reports of essays), which greatly differs from PERSONAGE's narrow dialogue domain.

Interestingly, we observe in Figure 8.11(c), 8.12(a) and 8.12(b) that the accuracy of the source models on the test data in the target domain is higher *without* any target training data at all (0% relative size), rather than with a small number of instances in the target domain. A possible explanation is that the feature duplication greatly increases the size of the feature space for almost no additional data, thus degrading learning performance.

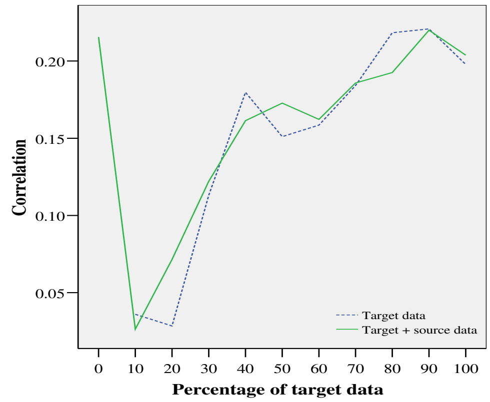
(a) Extraversion (M5' tree with LIWC features and observer reports).



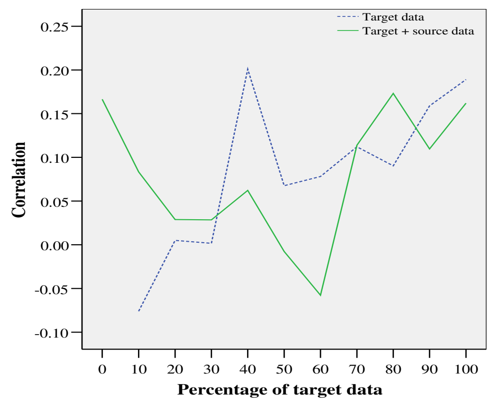
(b) Emotional stability (M5' tree with n-gram features and observer reports).

(c) Agreeableness (SVM with LIWC features and self-reports).

Figure 8.11: Correlation of the judges' ratings with the predictions of the best out-of-domain models for extraversion, emotional stability and agreeableness. Models are trained on target datasets of different sizes, with and without out-of-domain data.



(a) Conscientiousness (SVM with LIWC features and self-rerports).



(b) Openness to experience (M5' tree with n-gram features and self-reports).

Figure 8.12: Correlation of the judges' ratings with the predictions of the best out-of-domain models for conscientiousness and openness to experience. Models are trained on target datasets of different sizes, with and without out-of-domain data.

This exploratory analysis suggests that Daume's feature duplication is a promising technique for re-using data to improve personality modelling. However, our results suggest that the source domain should not differ too much from the target domain, i.e. personality perception from spoken language.

8.5 Summary

This chapter shows that it is possible to project personality over a continuous scale, by implementing the overgenerate and select paradigm in the PERSONAGE-OS data-driven generator. The generator's selection models perform significantly better than the random selection baseline. LIWC word category counts and generation decision features provide the best performance on unseen utterances, whereas n-gram features are more likely to overfit. This chapter also reveals that simple models such as linear regression models and regression trees predict personality as well or better than more complex models such as model trees and support vector machines with a non-linear kernel. While most previous work focuses on modelling error (apart from Paiva [2004]), we also evaluate the sampling error, showing that models with a large output variation are better suited to the utterance selection task. While we study both components of the error on our dataset, additional human evaluation would be beneficial to compare this generation method with the rule-based approach.

Selection models are better at predicting utterances expressing a mid-range personality, due to the normal distribution of the judges' ratings. To address this issue, we present a new approach combining the stochastic exploration of the parameter space with findings from psychology studies, by adding utterances generated using the rule-based method evaluated in Chapter 6 to the candidate utterance set. The addition of rule-based utterances conveying extreme personality significantly increases model performance for four traits out of five, while improving the coverage of the overgeneration phase.

This chapter also shows that models trained on general out-of-domain data outperform the random baseline in PERSONAGE's information presentation domain. They perform worse than models trained on in-domain data, however they can be re-used in different domains without additional data collection. Finally, we find that out-of-domain data can be used to enhance models trained on in-domain data, although our study is exploratory—future work requires further evaluation with a larger out-of-domain dataset.

The work of Isard et al. [2006] on personality generation seems to be based on the assumption that out-of-domain corpora can be used when there is no in-domain corpora available. While the out-of-domain corpora may work better if one had access to a larger corpus, or if the language models were blended with other data—as Isard et al. do with the Switchboard corpus—our results suggest that there is no data like in-domain data.

While the overgenerate and select paradigm has been implemented in the past

[Bangalore and Rambow, 2000, Langkilde-Geary, 2002, Walker et al., 2002, Isard et al., 2006], the computational time required to generate a large candidate utterance set makes real-time generation difficult with current technology. A recent line of work focuses on a new approach that does not require any overgeneration phase, by learning models that estimate the generation parameters directly from the desired target values in the variation space [Paiva and Evans, 2005]. Hence, the next chapter presents the first application of a *direct* data-driven generation technique for projecting personality, together with a large-scale evaluation of the perception of the personality conveyed over a continuous scale.

Chapter 9

Generation of Personality through Data-driven Parameter Estimation

Chapter 8 presents a first method for generating continuous personality variation, by using the overgenerate and select paradigm. While selection models trained on utterance ratings can select an utterance close to any value on the Big Five scales, the computational cost of generating a large candidate utterance set makes it impractical for real-time use. Paiva and Evans [2005] present a data-driven method for stylistic generation that does not require any overgeneration phase, which we extend in multiple ways. First, we focus on the control of the speaker's personality, rather than stylistic dimensions extracted from corpora. Second, we present a method for learning *parameter estimation models* predicting generation decisions directly from input personality scores, whereas Paiva and Evans' generator requires a search for the optimal generation decision over the model's input space. Third, we present an evaluation of the generated stylistic variation, according to human judges. Results show that the parameter estimation approach automatically learns to produce recognisable variation along the Big Five personality dimensions. The extension of the PERSONAGE generator with parameter estimation models is referred to as the PERSONAGE-PE trainable generator in the rest of this chapter.

The parameter estimation technique is described in Section 9.1, and a large-scale human perceptual evaluation is presented in Section 9.2. We show that PERSONAGE-PE can simultaneously hit scalar targets across multiple dimensions, using linear and non-linear parameter estimation models. Section 9.2 shows that the subjects accurately perceive the intended variation, and results are compared to the rule-based version of PERSONAGE presented in Chapter 6.

9.1 Methodology

The parameter estimation method consists of a development phase, described in this section, and a generation phase. The development phase consists of the following steps:

1. Use a base generator to produce multiple utterances by randomly varying its parameters (see Chapter 7);
2. Ask human subjects to evaluate the personality of each utterance;
3. Train statistical models predicting the parameter values from the personality ratings (see Section 9.1.2);
4. Select the best model for each parameter via cross-validation (see Section 9.1.4).

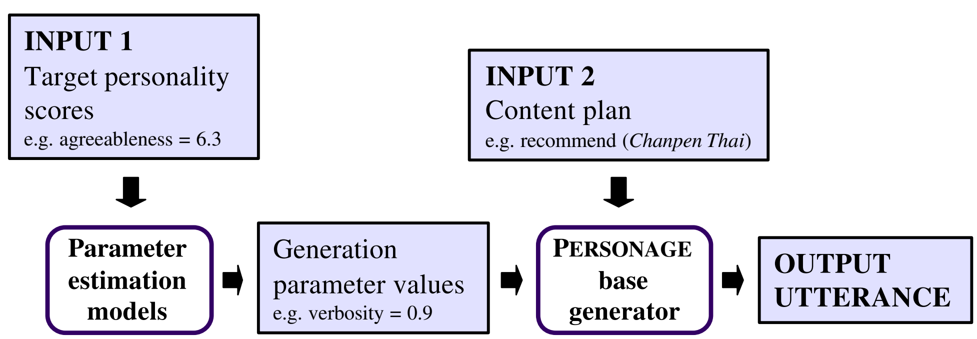


Figure 9.1: PERSONAGE-PE's parameter estimation framework.

PERSONAGE-PE is an extension of the rule-based PERSONAGE generator, in which the generation parameters in Tables 9.1 and 9.2 are controlled by parameter estimation models that estimate their parameter values from target personality scores. The generation phase is illustrated in Figure 9.1. Parameter estimation models are trained on the same data as the selection models in Chapter 8, i.e. personality ratings of 160 randomly generated utterances (320 for extraversion) obtained from the Ten-Item Personality Inventory [TIPI; Gosling et al., 2003], averaged over two judges (three for extraversion). However, the parameter estimation task requires a series of pre-processing steps, in order to ensure that the models' output is re-usable by the PERSONAGE base generator.

Continuous parameters	Description
Content planning:	
VERBOSITY	Control the number of propositions in the utterance
RESTATEMENTS	Paraphrase an existing proposition, e.g. *'Chanpen Thai has great service, it has fantastic waiters'*
REPETITIONS	Repeat an existing proposition
CONTENT POLARITY	Control the polarity of the propositions expressed, i.e. referring to negative or positive attributes
REPETITIONS POLARITY	Control the polarity of the restated propositions
CONCESSIONS	Emphasise one attribute over another, e.g. *'even if Chanpen Thai has great food, it has bad service'*
CONCESSIONS POLARITY	Determine whether positive or negative attributes are emphasised
POLARISATION	Control whether the expressed polarity is neutral or extreme
POSITIVE CONTENT FIRST	Determine whether positive propositions—including the claim—are uttered first
Syntactic template selection:	
SELF-REFERENCES	Control the number of first person pronouns
SYNTACTIC COMPLEXITY	Control the syntactic complexity (syntactic embedding)
TEMPLATE POLARITY	Control the connotation of the claim, i.e. whether positive or negative affect is expressed
Aggregation operations:	
PERIOD	Leave two propositions in their own sentences, e.g. *'Chanpen Thai has great service. It has nice decor.'*
RELATIVE CLAUSE	Aggregate propositions with a relative clause, e.g. *'Chanpen Thai, which has great service, has nice decor'*
WITH CUE WORD	Aggregate propositions using *with*, e.g. *'Chanpen Thai has great service, with nice decor'*
CONJUNCTION	Join two propositions using a conjunction, or a comma if more than two propositions
MERGE	Merge the subject and verb of two propositions, e.g. *'Chanpen Thai has great service and nice decor'*
ALSO CUE WORD	Join two propositions using *also*, e.g. *'Chanpen Thai has great service, also it has nice decor'*
CONTRAST - CUE WORD	Contrast two propositions using *while, but, however, on the other hand*, e.g. *'While Chanpen Thai has great service, it has bad decor'*, *'Chanpen Thai has great service, but it has bad decor'*
JUSTIFY - CUE WORD	Justify a proposition using *because, since, so*, e.g. *'Chanpen Thai is the best, because it has great service'*
CONCEDE - CUE WORD	Concede a proposition using *although, even if, but/though*, e.g. *'Although Chanpen Thai has great service, it has bad decor'*, *'Chanpen Thai has great service, but it has bad decor though'*
MERGE WITH COMMA	Restate a proposition by repeating only the object, e.g. *'Chanpen Thai has great service, nice waiters'*
OBJECT ELLIPSIS	Restate a proposition after replacing its object by an ellipsis, e.g. *'Chanpen Thai has . . . , it has great service'*
Pragmatic markers:	
SUBJECT IMPLICITNESS	Make the restaurant implicit by moving the attribute to the subject, e.g. *'the service is great'*
STUTTERING	Duplicate the first letters of a restaurant's name, e.g. *'Ch-ch-anpen Thai is the best'*
PRONOMINALISATION	Replace occurrences of the restaurant's name by pronouns
Lexical choice:	
LEXICON FREQUENCY	Control the average frequency of use of each content word, according to BNC frequency counts
LEXICON WORD LENGTH	Control the average number of letters of each content word
VERB STRENGTH	Control the strength of the verbs, e.g. *'I would suggest'* vs. *'I would recommend'*

Table 9.1: PERSONAGE's continuous generation parameters whose target values are learnt.

Binary parameters	Description
Content planning:	
REQUEST CONFIRMATION	Begin the utterance with a confirmation of the restaurant's name, e.g. *'did you say Chanpen Thai?'*
INITIAL REJECTION	Begin the utterance with a mild rejection, e.g. *'I'm not sure'*
COMPETENCE MITIGATION	Express the speaker's negative appraisal of the hearer's request, e.g. *'everybody knows that . . . '*
Pragmatic markers:	
NEGATION	Negate a verb by replacing its modifier by its antonym, e.g. *'Chanpen Thai doesn't have bad service'*
SOFTENER HEDGES	Insert syntactic elements (*sort of*, *kind of*, *somewhat*, *quite*, *around*, *rather*, *I think that*, *it seems that*, *it seems to me that*) to mitigate the strength of a proposition, e.g. *'Chanpen Thai has kind of great service'* or *'It seems to me that Chanpen Thai has rather great service'*
EMPHASISER HEDGES	Insert syntactic elements (*really*, *basically*, *actually*, *just*) to strengthen a proposition, e.g. *'Chanpen Thai has really great service'* or *'Basically, Chanpen Thai just has great service'*
ACKNOWLEDGMENTS	Insert an initial back-channel (*yeah*, *right*, *ok*, *I see*, *oh*, *well*), e.g. *'Well, Chanpen Thai has great service'*
FILLED PAUSES	Insert syntactic elements expressing hesitancy (*like*, *I mean*, *err*, *mmhm*, *you know*), e.g. *'I mean, Chanpen Thai has great service, you know'* or *'Err... Chanpen Thai has, like, great service'*
EXCLAMATION	Insert an exclamation mark, e.g. *'Chanpen Thai has great service!'*
EXPLETIVES	Insert a swear word, e.g. *'the service is damn great'*
NEAR EXPLETIVES	Insert a near-swear word, e.g. *'the service is darn great'*
TAG QUESTION	Insert a tag question, e.g. *'the service is great, isn't it?'*
IN-GROUP MARKER	Refer to the hearer as a member of the same social group, e.g. *pal*, *mate* and *buddy*

Table 9.2: PERSONAGE's binary generation parameters whose target values are learnt. Hedges, acknowledgments and filled pauses are learnt individually, e.g. *kind of* is modelled differently than *somewhat* in the SOFTENER HEDGES category.

9.1.1 Pre-processing steps

The initial dataset includes the generation decision features defined in Chapter 7 (Section 7.2.1) for each randomly generated utterance, together with the average judge ratings along each Big Five dimension. The following transformations are performed before the learning phase:

Reverse input and output: As parameter estimation models map from personality scores to generation parameters, the generation decisions are set as the dataset's output variables and the averaged personality ratings as the input features.

Predict parameters individually: A new dataset is created for each output variable—i.e. generation parameter—as the statistical models we use only predict one output. We thus make the simplifying assumption that PERSONAGE's generation parameters are independent.[1]

[1]While this assumption is violated by the internal constraints of PERSONAGE's generation process, Section 9.2 investigates the extent to which this violation affects the models' accuracy.

Map output variables into PERSONAGE's input space: As in Chapter 7, the actual generation decisions made for each utterance are recorded. These differ from the input parameters which are not always satisfiable depending on earlier decisions and the input content plan.[2] In order to ensure that the parameter estimation models' output is re-usable by the base generator, the generation decision space is mapped to PERSONAGE's input parameter space. The conversion is dependent on the type of generation parameter:

- **Continuous parameters:** Generation decision values are normalised over all random utterances, resulting in values between 0 and 1. E.g. a VERBOSITY parameter value of 1 indicates the utterance with the largest number of propositions in the utterance set.
- **Aggregation operation probabilities:** Frequency counts of aggregation operations realising a specific rhetorical relation are divided by the number of occurrences of the rhetorical relation in the utterance. This ratio is the maximum likelihood estimate of the conditional probability of the aggregation operation given the rhetorical relation. E.g. if out of four INFER relations in the utterance, only one is realised using the MERGE operation, the value for the INFER - MERGE parameter is .25 for that utterance.
- **Binary parameters:** No processing is required as generation decisions are already boolean. E.g. if an exclamation mark was inserted in the utterance, the EXCLAMATION parameter value is set to 1 rather than 0.

Feature selection: In order to only consider meaningful independent variables, personality traits that do not correlate with a generation parameter with a Pearson's correlation coefficient above .1 are removed from that parameter's dataset. This also has the effect of removing parameters that do not correlate strongly with any trait, which are set to a constant default value at generation time. Thus, parameter estimation models contain a maximum of five features (one for each Big Five trait), but they contain less than five features in a number of cases.

Once the data is partitioned into datasets mapping the relevant personality dimensions (the features) to each generation parameter (the dependent variable), it can be used to train parameter estimation models predicting the most appropriate parameter value given target personality scores.

[2]E.g. the CONCESSIONS decision value is the actual number of CONCEDE rhetorical relations produced, rather than the input probability between 0 and 1.

9.1.2 Statistical learning algorithms

Parameters are estimated using either regression or classification models, depending on whether they are continuous (e.g. VERBOSITY) or binary (e.g. EXCLAMATION). As in the previous chapter, we compare various learning algorithms using the Weka toolbox [Witten and Frank, 2005]. We use the same models as for the personality recognition task in Chapter 3.

Continuous parameters in Table 9.1 are modelled with a linear regression model (LR), an M5' model tree (M5), and a model based on support vector machines with a linear kernel (SVM). As regression models can extrapolate beyond the $[0, 1]$ interval, the output parameter values are truncated if needed—at generation time—before being sent to the base generator. Regression models are evaluated using the correlation between the model's predictions and the actual parameter values in the test data.

Binary parameters in Table 9.2 are modelled using classifiers that predict whether the parameter should be *enabled* or *disabled*. We test a Naive Bayes classifier (NB), a C4.5 decision tree (J48), a nearest neighbour classifier using one neighbour (NN), the Ripper rule-based learner (JRIP), the AdaBoost boosting algorithm (ADA) and a support vector machines classifier with a linear kernel (SVM). Unless specified, the learning algorithms use Weka's default parameter values. Classification models are evaluated using the F-measure of the *enabled* class, which is defined as:

$$\text{F-measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The F-measure is thus the weighted harmonic mean of the recall and the precision of the *enabled* class, with the recall being the ratio of correct *enabled* predictions over the total number of *enabled* parameter values in the data, and the precision being the ratio of correct *enabled* predictions over the total number of *enabled* predictions made by the model. As opposed to the classification accuracy metric, the F-measure favours models that do not always output *disabled* (recall $= 0$) despite the small proportion of instances labelled as *enabled* in the data.

9.1.3 Qualitative model analysis

Figures 9.2, 9.3 and 9.4 show models learned for the EXCLAMATION (binary), STUTTERING (continuous) and CONTENT POLARITY (continuous) parameters. See Tables 9.1 and 9.2 for a description of each parameter. The figures illustrate how the models predict generation parameters from input personality scores.

```
Weak classifier: decision stump

Condition                  Class                        Weight
---------                  -----                        ------
if extraversion  >  6.42   then enabled else disabled   1.81
if extraversion  >  4.42   then enabled else disabled   0.38
if extraversion  <= 6.58   then enabled else disabled   0.22
if extraversion  >  4.71   then enabled else disabled   0.28
if agreeableness >  5.13   then enabled else disabled   0.42
if extraversion  <= 6.58   then enabled else disabled   0.14
if extraversion  >  4.79   then enabled else disabled   0.19
if extraversion  <= 6.58   then enabled else disabled   0.17
```

Figure 9.2: AdaBoost model predicting the EXCLAMATION parameter. Given input trait values, the model outputs the class yielding the largest sum of weights for the rules returning that class.

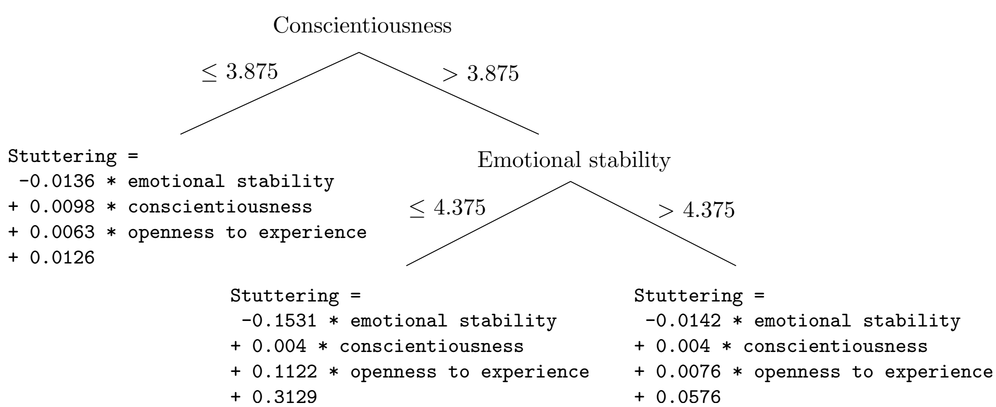


Figure 9.3: M5' model tree predicting the STUTTERING parameter.

Given input trait values, the AdaBoost model in Figure 9.2 outputs the class yielding the largest sum of weights for the rules returning that class. For example, the first rule of the EXCLAMATION model shows that an extraversion score above 6.42 out of 7 would increase the weight of the *enabled* class by 1.81. The fifth rule indicates that a target agreeableness above 5.13 would further increase the weight by .42. Figure 9.2 also illustrates how personality traits that do not have an effect on the parameter are removed, i.e. extraversion and agreeableness are the traits that affect the use of exclamation marks. The STUTTERING model tree in Figure 9.3 lets us calculate that a low emotional stability (1.0) together with a neutral conscientiousness and openness to experience (4.0) yield a parameter value of .62 (see bottom-left linear model), whereas a neutral emotional stability decreases the value down to .17. The linear model in Figure 9.4 shows that agreeableness has a strong effect on the CONTENT POLARITY parameter (.97 weight), but emotional stability, conscientiousness and openness to experience also influence the parameter value.

```
Kernel used: Linear Kernel: K(x,y) = <x,y>

(normalized) Content polarity  =
  -0.102 * (normalized) emotional stability
 + 0.970 * (normalized) agreeableness
 - 0.110 * (normalized) conscientiousness
 + 0.013 * (normalized) openness to experience
 + 0.054
```

Figure 9.4: SVM model with a linear kernel predicting the CONTENT POLARITY parameter.

9.1.4 Model selection

The final step of the development phase is to identify the best performing model(s) for each generation parameter via a 10-fold cross-validation. For continuous parameters, Table 9.3 evaluates modelling accuracy by comparing the correlations between the model's predictions and the actual parameter values in the test folds. Table 9.4 reports results for binary parameter classifiers, by comparing the F-measures of the *enabled* class. Best performing models are identified in bold for each parameter; parameters that do not correlate with any trait or that produce a poor modelling accuracy are omitted.

The CONTENT POLARITY parameter is modelled the most accurately, with the SVM model shown in Figure 9.4 producing a correlation of .47 with the true parameter values in Table 9.3. Models of the PERIOD aggregation operation also perform well, with a linear regression model yielding a correlation of .36 when realising a justification, and .27 when contrasting two propositions. The SYNTACTIC COMPLEXITY and VERBOSITY parameters are also modelled successfully, with correlations of .33 and .26 using a model tree. The model tree controlling the STUTTERING pa-

Continuous parameters	LR	M5	SVM
Content planning:			
Verbosity	0.24	**0.26**	0.21
Restatements	0.14	**0.14**	0.04
Repetitions	**0.13**	0.13	0.08
Content polarity	0.46	0.46	**0.47**
Repetition polarity	0.02	**0.15**	0.06
Concessions	**0.23**	0.23	0.12
Concession polarity	-0.01	**0.16**	0.07
Polarisation	0.20	**0.21**	0.20
Syntactic template selection:			
Syntactic complexity	0.10	**0.33**	0.26
Template polarity	0.04	0.04	**0.05**
Aggregation operations:			
Infer - with cue word	0.03	**0.03**	0.01
Infer - also cue word	**0.10**	0.10	0.06
Justify - since cue word	0.03	**0.07**	0.05
Justify - so cue word	0.07	**0.07**	0.04
Justify - period	**0.36**	0.35	0.21
Contrast - period	**0.27**	0.26	0.26
Restate - merge with comma	0.18	**0.18**	0.09
Concede - although cue word	**0.08**	0.08	0.05
Concede - even if cue word	0.05	**0.05**	0.03
Pragmatic markers:			
Subject implicitness	**0.13**	0.13	0.04
Stuttering	0.16	**0.23**	0.17
Pronominalisation	**0.22**	0.20	0.17
Lexical choice:			
Lexicon frequency	0.21	**0.21**	0.19
Lexicon word length	**0.18**	0.18	0.15

Table 9.3: Pearson's correlation coefficient between parameter model predictions and continuous parameter values, for different regression models. Parameters that do not correlate with any trait are omitted. Results are averaged over a 10-fold cross-validation.

rameter illustrated in Figure 9.3 produces a correlation of .23. Concerning binary parameters, Table 9.4 shows that the Naive Bayes classifier is generally the most accurate, with F-measures of .40 for the IN-GROUP MARKER parameter, and .32 for both the insertion of filled pauses (*err*) and tag questions. The AdaBoost learning algorithm performs best for predicting the EXCLAMATION parameter, with an F-measure of .38 for the model shown in Figure 9.2.

9.1.5 Generation phase

Once the best parameter estimation models have been identified, they can be used to generate utterances expressing any combination of personality target scores. The generation phase consists of the following steps:

Binary parameters	**NB**	**J48**	**NN**	**JRIP**	**ADA**	**SVM**
Content planning:						
REQUEST CONFIRMATION	0.00	0.00	**0.07**	0.05	0.04	0.04
Pragmatic markers:						
SOFTENER HEDGES						
kind of	0.00	0.00	**0.16**	0.13	0.11	0.10
rather	0.00	0.00	**0.02**	0.02	0.01	0.01
quite	**0.14**	0.08	0.09	0.09	0.07	0.06
EMPHASISER HEDGES						
basically	0.00	0.00	**0.02**	0.01	0.01	0.01
ACKNOWLEDGMENTS						
yeah	0.00	0.00	**0.04**	0.04	0.03	0.03
ok	**0.13**	0.07	0.06	0.05	0.05	0.05
FILLED PAUSES						
err	**0.32**	0.20	0.24	0.24	0.22	0.19
EXCLAMATION	0.23	0.34	0.36	0.37	**0.38**	0.34
EXPLETIVES	**0.27**	0.18	0.24	0.20	0.17	0.15
IN-GROUP MARKER	**0.40**	0.31	0.31	0.26	0.24	0.21
TAG QUESTION	**0.32**	0.21	0.21	0.17	0.15	0.13

Table 9.4: F-measure of the *enabled* class for classification models of binary parameters. Parameters that do not correlate with any trait are omitted. Results are averaged over a 10-fold cross-validation.

1. Use the best performing models to predict parameter values from the desired personality scores;
2. Generate the output utterance using the predicted parameter values.

The following section presents an evaluation using a large number of naive human judges to rate the perceived personality and naturalness of PERSONAGE-PE's output.

9.2 Large-scale evaluation

Whereas Section 9.1.4 evaluates the accuracy of parameter estimation models to predict parameter values from test data, it does not address many questions related to the generalisation of these results:

Q1: Is the personality conveyed by models trained on ratings from a small number of expert judges recognised by a larger sample of naive judges? (Section 9.2.2)

Q2: Can a *combination* of multiple traits within a single utterance be detected by naive judges? (Section 9.2.2)

Q3: How does the parameter estimation method compare to a psychologically-informed rule-based generator for projecting extreme personality? (Section 9.2.3)

Q4: Do direct generation models successfully project personality *continuously*, i.e. can naive judges detect fine-grained personality variation? (Section 9.2.4)

Q5: Does the parameter estimation method produce natural utterances? (Section 9.2.6)

We present a large-scale evaluation of PERSONAGE-PE, which addresses these open questions in the sections below.

9.2.1 Evaluation method

Given the best performing model for each generation parameter, we generate 5 utterances for each of 10 input content plans. Each utterance targets an extreme value for two traits (either 1 or 7 out of 7) and neutral values for the remaining three traits (4 out of 7). The goal is for each utterance to project *multiple* traits on a *continuous* scale. In order to generate a range of alternatives for each input content plan, all target scores are randomised around their initial value according to a normal probability distribution with a standard deviation of 10% of the full scale (see Figure 9.5).

Subjects were 24 native English speakers (12 male and 12 female graduate students from a range of disciplines from both the U.K. and the U.S.). Subjects evaluate the personality of each utterance using a subset of the Ten-Item Personality Inventory detailed in Figure 6.2 in Chapter 6 [TIPI; Gosling et al., 2003]. To limit the experiment's duration, only the two traits with extreme target values are evaluated for each utterance. Each utterance was also evaluated for its naturalness, defined as whether it could have been said by a human. Subjects thus answered 5 questions for 50 utterances, two from the TIPI for each extreme trait and one about naturalness (250 judgements in total per subject). Subjects were not told that the utterances were intended to manifest extreme trait values.

Table 9.5 shows several sample outputs and the mean personality ratings from the human judges for two content plans. For example, utterance 1.a projects a high extraversion through the insertion of an exclamation mark based on the model in

#	Trait and target		Rating	Nat	Output utterance
1.a	Extraversion	high 6.90	4.42	4.79	Radio Perfecto's price is 25 dollars but Les Routiers provides adequate food. I imagine they're alright!
	Agreeableness	high 7.00	4.94		
1.b	Emotional stability	high 7.00	5.35	5.04	Let's see, Les Routiers and Radio Perfecto... You would probably appreciate them. Radio Perfecto is in the East Village with kind of acceptable food. Les Routiers is located in Manhattan. Its price is 41 dollars.
	Conscien-tiousness	high 6.21	5.21		
1.c	Agreeableness	low 1.28	4.04	4.58	Les Routiers and Radio Perfecto are outstanding. Radio Perfecto's price is 25 dollars, isn't it? The food is acceptable, also it's located in the East Village. Err... although Les Routiers offers adequate food, it's damn costly.
	Openness to experience	high 7.00	4.52		
2.a	Extraversion	low 1.92	3.65	3.21	Err... you would probably appreciate Trattoria Rustica, wouldn't you? It's in Manhattan, also it's an italian restaurant. It offers poor ambience, also it's quite costly.
	Agreeableness	low 1.73	4.02		
2.b	Emotional stability	low 1.97	4.13	4.50	Trattoria Rustica isn't as bad as the others. Err... even if it's costly, it offers kind of adequate food, alright? It's an italian place.
	Openness to experience	low 1.00	3.85		
2.c	Agreeableness	low 1.00	3.27	3.58	Trattoria Rustica is the only eating place that is any good. Err... it's located in Manhattan. This restaurant is an italian place with poor ambience. It's bloody costly, even if this eating house has friendly waiters you see?
	Openness to experience	low 1.33	3.94		

Table 9.5: Example outputs from PERSONAGE-PE for a comparison (#1) and a recommendation (#2), with the average judges' personality (*Rating*) and naturalness (*Nat*) scores. Ratings are on a scale from 1 to 7, with 1 = very low (e.g. introvert) and 7 = very high (e.g. extravert).

Figure 9.2, whereas utterance 2.a conveys introversion by beginning with the filled pause *err*. The same utterance also projects a low agreeableness by focusing on negative propositions, through a low CONTENT POLARITY parameter value produced by the model in Figure 9.4. The 50 utterances generated for the evaluation are listed in Appendix C.

9.2.2 Evaluation results

Table 9.6 shows that extraversion is the dimension modelled the most accurately by the parameter estimation models, producing a .45 correlation between the target extraversion and the subjects' ratings ($p < .01$). Emotional stability, agreeableness and openness to experience ratings also correlate strongly with the target scores, with correlations of .39, .36 and .17 respectively ($p < .01$). Additionally, Table 9.6 shows that the magnitude of the correlation increases when considering the perception of a hypothetical average subject—i.e. smoothing individual variation by averaging the ratings over all 24 judges—producing a correlation r_{avg} up to .80 for extraversion. These correlations are unexpectedly high; in corpus analyses, significant correlations as low as .05 to .15 are typically observed between averaged personality ratings and linguistic markers [Pennebaker and King, 1999, Mehl et al., 2006].

Trait	r	r_{avg}	e
Extraversion	.45 •	.80 •	1.89
Emotional stability	.39 •	.64 •	2.14
Agreeableness	.36 •	.68 •	2.38
Conscientiousness	-.01	-.02	2.79
Openness to experience	.17 •	.41 •	2.51

• statistically significant correlation
$p < .05$, • $p = .07$ (two-tailed)

Table 9.6: Pearson's correlation coefficient r and mean absolute error e between the target personality scores and the 480 judges' ratings (20 ratings per trait for 24 judges); r_{avg} is the correlation between the target scores and the 20 ratings averaged over all judges.

Conscientiousness is the only dimension whose ratings do not correlate with the target scores. The comparison with rule-based results in Section 9.2.3 suggests that this is not because conscientiousness cannot be exhibited in our domain or manifested in a single utterance, so perhaps this arises from differing perceptions of conscientiousness between the expert and naive judges.

Table 9.6 shows that the mean absolute error varies between 1.89 and 2.79 on a scale from 1 to 7. Such large errors result from the decision to ask judges to answer just the TIPI questions for the two traits that were the extreme targets (as described in Section 9.2.1), because the judges tend to use the whole scale, with approximately normally distributed ratings. This means that although the judges make distinctions leading to high correlations, the averaged ratings result in a compressed scale. This explains the large correlations despite the magnitude of the absolute error.

Table 9.7 reports results evaluating whether utterances targeting the extremes

Trait	Low	High
Extraversion	3.69	5.06 •
Emotional stability	3.75	4.75 •
Agreeableness	3.42	4.33 •
Conscientiousness	4.16	4.15
Openness to experience	3.71	4.06 •

• statistically significant difference
$p \leq .001$ (two-tailed)

Table 9.7: Average personality ratings for the utterances generated with the low and high target values for each trait on a scale from 1 to 7.

of a trait are perceived differently. T-tests show that the ratings differ significantly for all traits but conscientiousness ($p \leq .001$). Thus parameter estimation models can be used in applications that only require discrete binary variation.

It is important to emphasise that generation parameter values were predicted based on five target personality scores. Thus, the results show that *individual* traits are perceived even when utterances project other traits as well, confirming that the Big Five theory models independent dimensions and thus provides a useful framework for modelling variation in language.

9.2.3 Comparison with rule-based generation

PERSONAGE-RB is a rule-based personality generator based on handcrafted parameter settings derived from psychological studies. Chapter 6 shows that it generates utterances that are perceptibly different along all Big Five dimensions. Table 9.8 compares the mean ratings of the utterances generated by PERSONAGE-PE with ratings of 20 utterances generated with PERSONAGE-RB's predefined parameter settings for each extreme of each Big Five scale (40 for extraversion, resulting in 240 rule-based utterances in total). Table 9.8 shows that the handcrafted parameter settings project a significantly more extreme personality for 6 traits out of 10. However, the parameter estimation models have not been shown to perform significantly worse than the rule-based generator for neuroticism, disagreeableness, unconscientiousness and openness to experience. In spite of these findings, parameter estimation models are promising as (1) they are able to target any combination of traits over the full range of the Big Five scales; (2) they do not benefit from psychological knowledge, i.e. they are trained on randomly generated utterances; (3) presumably the accuracy of the parameter estimation models could be improved with a larger number of expert judges and random utterances at development time.

Method	**Rule-based**		**Param models**	
Trait	**Low**	**High**	**Low**	**High**
Extraversion	2.96	5.98	3.69 ∘	5.06 ∘
Emotional stability	3.29	5.96	3.75	4.75 ∘
Agreeableness	3.41	5.66	3.42	4.33 ∘
Conscientiousness	3.71	5.53	4.16	4.15 ∘
Openness to experience	2.89	4.21	3.71 ∘	4.06

•,∘ significant increase or decrease of the variation range over the average rule-based ratings ($p < .05$, two-tailed)

Table 9.8: Comparison between the ratings of PERSONAGE-PE's utterances with extreme target values (*Param models*) and the expert judges' ratings for utterances generated using PERSONAGE-RB in Chapter 6 (*Rule-based*).

9.2.4 Perception of fine-grained variation

The previous section shows that parameter estimation models generate utterances perceived as matching the target personality traits *in the large*, i.e. the judges discriminate between utterances with very different target scores. This section focuses on the modelling of personality *in the small*—a much harder task—by evaluating whether PERSONAGE-PE accurately projects finer-grained variation, e.g. within one unit on the 1... 7 scale.

We mention in Section 9.2.1 that the target scores used in the evaluation experiment were randomised according to a normal distribution around 1 or 7, with a standard deviation of 10% of the full scale (.60). Figure 9.5 shows the distribution of the target scores for emotional stability. The data can therefore be partitioned into groups of utterances projecting the same extreme traits with only small variations of the target score.[3] This section evaluates whether the judges perceive the small differences within each group of extreme utterances, by computing the correlation between the target scores and the judges' ratings over each group.

Table 9.9 shows that the judges detect the small variation along the emotional stability scale, with correlations of .19 for the neurotic group and .33 for the emotionally stable group ($p < .01$). The ratings of the average user correlate even more strongly ($r_{avg} = .46$ and $.55$ respectively, marginally significant). These results con-

[3]Because target scores are truncated to fit PERSONAGE-PE's input range (between 1 and 7), approximately half of the values in each group are either 1.0 or 7.0.

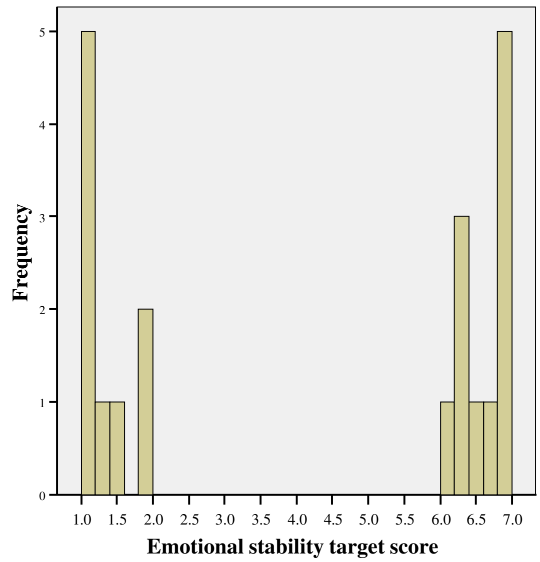


Figure 9.5: Distribution of the 20 emotional stability target scores, normally distributed over both extremes with a standard deviation of 10% of the full scale.

	Correlation r		Correlation r_{avg}		Range	
Trait	**Low**	**High**	**Low**	**High**	**Low**	**High**
Extraversion	.00	.02	.01	.05	.92	1.31
Emotional stability	.19 •	.33 •	.46	.55 •	.97	.98
Agreeableness	.09	-.07	.20	-.17	1.13	.70
Conscientiousness	.02	-.18	.03	-.32	.97	.79
Openness to experience	.03	.18 •	.08	.46	.84	1.52

• statistically significant correlation
$p < .05$, • $p = .08$ (two-tailed)

Table 9.9: Pearson's correlation coefficient between the target personality scores and individual ratings (r) and averaged ratings (r_{avg}) for each group of extreme targets, as well as the target score range.

firm the high granularity of the parameter estimation models for that trait, as these correlations are observed over ranges smaller than 1.0 (respectively $[1.00, 1.97]$ and $[6.02, 7.00]$ for each group). The variation is also perceived for utterances projecting a high openness to experience ($r = .18$, $r_{avg} = .46$). Although ratings of utterances conveying a low agreeableness also correlate positively with the target scores ($r = .09$, $r_{avg} = .20$), additional experiments are required to show whether this result is significant ($p = .19$). The low correlation observed for other traits—e.g. extraversion—shows that the high accuracy reported in Section 9.2.2 is due to the successful modelling of large variations between each end of the scale, rather than the small-scale variations within one side of the dimension.

9.2.5 Inter-rater agreement

The level of agreement between judges reflects the difficulty of the personality recognition task for humans, thus providing an upper bound on the performance to be expected from a model trained on human data. Table 9.10 reports the inter-rater correlation over all personality ratings, averaged over the 276 pairs of judges. The judges agree modestly, with correlations ranging from .17 (openness to experience) to .41 (emotional stability). Although this agreement is lower than the one reported in Chapter 6 on utterances generated from predefined parameter settings, our experiment involves a much larger sample of *naive* judges. As personality perception is a non-trivial task for humans, naive judges are less likely to perceive personality in the same way as trained experts, and less likely to be consistent in their judgements. Interestingly, these correlations are comparable to the correlations between the judges' ratings and the target personality scores reported in Table 9.6 (apart from conscientiousness). This suggests that the parameter estimation models predict the ratings of the generated utterances as well as individual judges predict each others' ratings.

Trait	$\bar{r}_{inter}$	σ_{inter}
Extraversion	.33	.22
Emotional stability	.41	.17
Agreeableness	.28	.23
Conscientiousness	.34	.18
Openness to experience	.17	.25
All	.34	.10

Table 9.10: Average and standard deviation of the inter-rater correlations over the 276 pairs of judges.

9.2.6 Naturalness evaluation

Method	**Rule-based**	**Random**	**Param models**
Judges	Expert	Expert	Naive
Naturalness	4.59	4.38	3.98

Table 9.11: Average naturalness ratings for utterances generated using (1) the rule-based generator, (2) random parameters and (3) parameter estimation models (*Param models*). The means differ significantly at the $p < .05$ level (two-tailed independent sample t-test).

The naive judges also evaluated the naturalness of PERSONAGE-PE's outputs.

Table 9.11 shows that the average naturalness is 3.98 out of 7, which is significantly lower ($p < .05$) than the naturalness of rule-based and randomly generated utterances reported in Chapters 6 and 7. It is not clear what these differences are due to. It is possible that the differences arise from judgements of utterances targeting multiple traits, or that the naive judges are more critical. Figure 9.7 shows that the naturalness ratings tend to increase over time, probably because the judges become more familiar with the type of variation produced by the generator. As the expert judges rated considerably more utterances than the naive judges, this could explain the higher naturalness ratings.

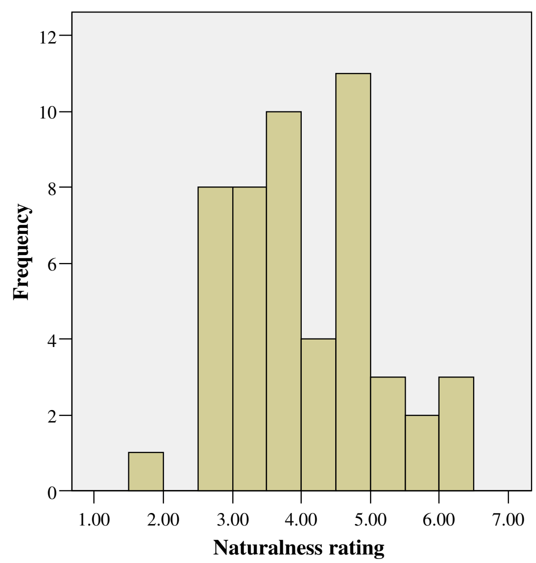
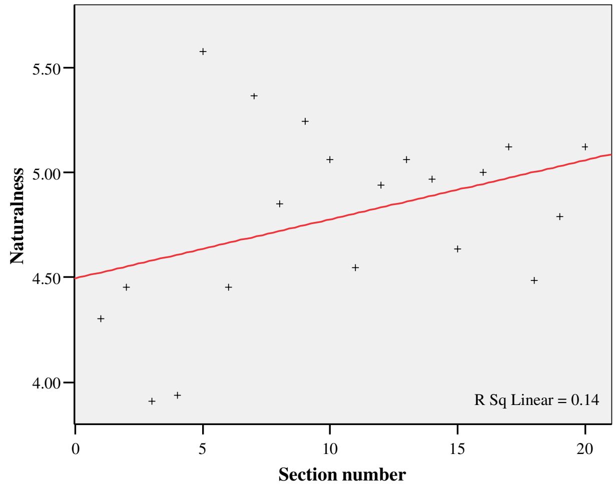


Figure 9.6: Distribution of naturalness ratings over the 50 utterances, averaged over all 24 judges. The mean naturalness rating is 3.98, with a standard deviation of 1.07.

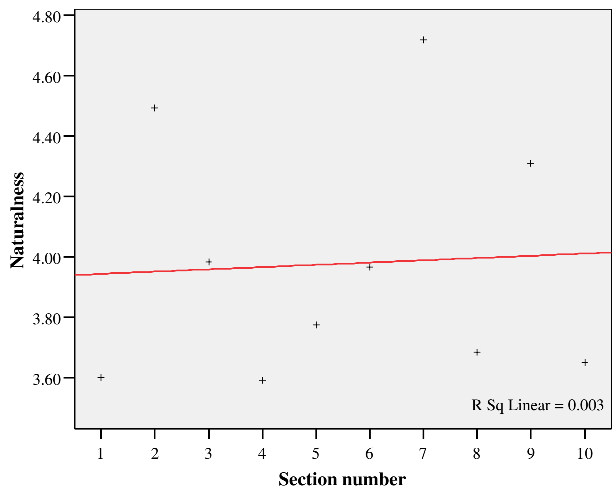
Figure 9.6 shows that the distribution of the naturalness ratings is almost normally distributed, with only one utterance out of 50 rated below 2.5 out of 7 on average.

9.2.7 Socio-cultural analysis

Because of the large number of judges involved in the evaluation, we can assess whether the ratings are affected by some of the judges' characteristics. In order to evaluate whether gender has an influence on the ratings, we compare personality ratings between the 12 male and 12 female judges involved in the experiment.



(a) Expert judges' ratings of the rule-based and random utterances from Chapters 6 and 7 (220 ratings per judge over 20 sections).



(b) Naive judges' ratings of PERSONAGE-PE's utterances (50 ratings per judge over 10 sections)

Figure 9.7: Evolution of the naturalness ratings over time, interpolated by a least-squares linear regression. Ratings are averaged over all judges within each section, and sections are ordered chronologically according to the online questionnaire.

Table 9.12 shows that female judges tend to rate utterances as more emotionally stable, as well as more natural ($p < .001$).

We also assess whether the judges' cultural background affects the ratings, by comparing results obtained with the 19 judges from the United Kingdom, to those obtained with the 4 judges from the United States.[4] This is motivated by the fact

[4]We ignore one judge from continental Europe.

Variable	Male	Female	p	UK	USA	p
Extraversion	4.68	4.75	.52	4.70	4.80	.54
Emotional stability	4.11	4.49 •	.00	4.24	4.53 •	.09
Agreeableness	3.87	3.88	.97	3.86	3.92	.69
Conscientiousness	4.11	4.19	.55	4.09	4.44 •	.05
Openness to experience	3.90	3.80	.36	3.88	3.63 •	.06
Naturalness	3.75	4.20 •	.00	3.86	4.60 •	.00
Correlation with target	.28	.34	.16	.32	.30	.64

• statistically significant difference
$p < .05$, • $p < .10$ (two-tailed)

Table 9.12: Comparison of personality ratings, naturalness ratings and correlation with target scores for different groups of participants, i.e. between male and female subjects, as well as between subjects from the United Kingdom (*UK*) and the United States (*USA*). Column p reports the level of significance of the difference of the means according to an independent samples t-test (two-tailed).

that most studies from which PERSONAGE's parameters are derived focus on American English, they might thus be perceived differently by British English speakers. However, Table 9.12 does not show any difference of correlation with the target scores between both groups. Results show that American judges tend to rate utterances as projecting more emotional stability, conscientiousness and openness to experience ($p < .10$). They also rate utterances as significantly more natural, with an average rating of 4.6 out of 7 ($p < .001$).

Table 9.12 also evaluates whether some groups of judges are better at detecting the personality cues produced by the parameter estimation models, however no significant results are found. Overall, these findings are difficult to interpret as they are likely to reflect complex socio-cultural phenomena. While their study is beyond the scope of this work, these results show that there is no absolute gold standard in personality perception, and that an ideal generator should take the user's background into account when projecting personality.

9.3 Discussion and summary

This chapter proposes a new method for generating linguistic variation projecting multiple personality traits continuously, by combining and extending previous research in statistical natural language generation [Langkilde-Geary, 2002, Walker et al., 2002, Paiva and Evans, 2005, Mairesse and Walker, 2007]. This method learns parameter estimation models trained on human judgements to predict the optimal generation decisions given target personality scores, without any overgeneration phase. The parameter estimation approach is implemented in the PERSONAGE-PE trainable generator.

While handcrafted rule-based approaches such as the one presented in Chapter 6 are limited to variation along a small number of discrete points [Hovy, 1988, Walker et al., 1997, Cassell and Bickmore, 2003], parameter estimation models learn to predict parameter values for any arbitrary value on the personality scales. Additionally, this data-driven approach can be applied to any other dimension that is meaningful to human judges. It also provides an elegant way to project multiple dimensions simultaneously, by including the relevant dimensions as features of the parameter estimation models' training data.

The overgenerate and select method presented in Chapter 8 is another data-driven approach to linguistic style variation. However, it requires the generation of a large number of utterances at runtime, which makes it inappropriate for real-time dialogue. Paiva and Evans [2005] also present a technique that does not overgenerate (detailed in Section 2.4.4.2 in Chapter 2), by training linear models predicting stylistic factors from generation decisions. The generation decisions yielding the desired stylistic scores are then estimated by searching over the models' input space. This chapter's parameter estimation approach does *not* require any search phase, as the models predict the generation decisions directly from the target variation dimensions. This technique is therefore beneficial for real-time generation.

This chapter also presents the first human perceptual evaluation of a data-driven stylistic variation method, showing that the perceived personality of PERSONAGE-PE's utterances correlates significantly with the target scores, except for the conscientiousness dimension. We also find that the parameter estimation models perform only slightly worse when projecting extreme traits than the PERSONAGE-RB rule-based generator in the same domain. These findings are promising as (1) parameter estimation models are able to target any scalar combination of the Big Five traits; (2) they do not benefit from psychological knowledge; and (3) their accuracy could be improved with a larger sample of expert judges and random utterances at development time. A larger number of judges would also smooth out rating inconsistencies and individual differences in personality perception, thus allowing the direct modelling of laypeople's perceptions by removing the need for expert judges.

While PERSONAGE's parameters were suggested by psychological studies, some of them are not modelled successfully by the parameter estimation approach, and thus were omitted from Tables 9.3 and 9.4. This could be due to the relatively small development dataset size (160 utterances to optimise 67 parameters), or to the implementation of specific parameters. Although the parameter-independence assumption mentioned in Section 9.1.1 could also be responsible for the poor accuracy of some models, it could possibly be resolved by training statistical models that simultaneously predict multiple dependent variables. However, we are not aware of any state of the art implementation of such learning algorithms,[5] and their use could further aggravate data sparsity problems. Despite these issues, the results presented in this chapter suggest that a limited number of parameter estimation models can successfully project multiple personality traits.

By building on results of Chapters 6 and 8, we provide further evidence that personality can be expressed in the information presentation domain. Future work should address why Chapter 6's rule-based method performs better for projecting extreme personality while producing more natural utterances, and why the parameter estimation method fails to project conscientiousness correctly.

[5]Although neural networks can predict multiple outputs simultaneously, they did not perform well on our single-output task and were not further investigated.

Chapter 10

Discussion and Conclusion

This thesis has investigated techniques for modelling individual differences in the context of dialogue system applications, such as spoken dialogue systems, interactive drama systems and intelligent tutoring systems. We model these differences through the concept of *personality*, by modelling the personality of the user as well as the personality conveyed by the system. Each method has been evaluated in detail, by testing (1) whether the personality of unseen speakers is successfully recognised and (2) whether human judges recognise the intended personality of the system.

In this chapter, Section 10.1 summarises the contributions of this research in terms of the research questions raised in Chapter 1. Then, Section 10.2 discusses the generalisation of this work to a new application domain: the generation of personality in interactive narrative systems. Finally, Section 10.3 reviews some limitations of this research, together with possible areas of future work.

10.1 Contributions of this thesis

This research has investigated two hypotheses stated in Chapter 1:

Hypothesis 1 *Statistical models can learn to predict the personality of unseen individuals from conversational data, on a continuous scale.*

Hypothesis 2 *The personality conveyed by generated utterances can be controlled by a computational model, which can be either derived from psycholinguistic knowledge or learnt from personality-annotated linguistic data.*

This thesis has provided positive evidence for both hypotheses, by addressing a number of research questions. We summarise the contributions of this work by reviewing them together with some answers suggested by our results.

Personality recognition:

Can statistical models trained on spoken language annotated with judgements of personality successfully predict the personality of unseen individuals?

Results presented in Chapter 3 show that personality classification and ranking models trained on personality-annotated spoken language samples significantly outperform our baseline for each Big Five trait. Binary classification models trained on observer reports of conversations produce between 65% and 74% correct classifications when tested on unseen subjects (see summary in Table 3.26). Concerning continuous personality modelling—which is a more difficult task—we find that the best regression model for extraversion produces a mean error that is 24% lower than the error made by a constant baseline (correlation of .54). Ranking models all outperform the random ranking baseline, by correctly ordering between 61% and 74% of all pairs of unseen subjects over all Big Five traits. However, accuracies vary depending on the personality trait, the source of language and the type of personality being modelled (see below).

How does the personality assessment method—e.g. using the speaker's own judgement or observer reports—affects modelling accuracy?

While psychology studies address this question by comparing correlations between personality reports and linguistic markers over the full dataset [Mehl et al., 2006], this thesis compares models trained on self-reports and observer reports by evaluating them on unseen data. A clear result that emerges is that observed extraversion is the easiest trait to model from conversations, as Table 3.26 shows that extraversion models perform best for all recognition tasks. This confirms the assumption that extraversion is the trait that is the most easily perceived through spoken language. Interestingly, we find that agreeableness is the hardest trait to model using the same dataset, producing the worst accuracies for all three recognition tasks. While models of self-reports of personality trained on conversation extracts never outperform the baseline (possibly due to the small dataset size), models trained on stream-of-consciousness essays all perform significantly better than the baseline, although the magnitude of the improvement is modest. Models of openness to experience perform best for the three recognition tasks, with 63% correct classifications on unseen essays, 7% improvement over the regression baseline, and 61% correct rankings between all pairs of essays in the test set. Models of other traits perform only slightly better than the baseline, suggesting that the author's personality is generally difficult to detect through his or her writings.

What linguistic features are the most useful for personality recognition?

In Chapter 3, we systematically compare modelling accuracies using feature sets characterising the type of utterance produced, syntactic and semantic word categories used in psychology studies, and prosody. While psychology studies already investigate this question [Pennebaker and King, 1999, Mehl et al., 2006], such correlational analyses do not reflect the accuracy of non-linear models, and they only focus on content-analysis category features. Results summarised in Table 3.26 show that the LIWC content-analysis categories emerge as the best feature set for modelling self-reports of personality from essays, as they perform best for all traits on both the classification and ranking tasks. Concerning observer reports of conversations, results are less clear. The full feature set produces the best classification models for four traits out of five. The best regression model for extraversion and conscientiousness are obtained with the LIWC features, whereas emotional stability is modelled best using only prosodic features. Interestingly, the ranking model for extraversion—the best ranking model out of all Big Five traits—uses only prosodic features as well, suggesting that non-verbal cues can be more important than linguistic cues for conveying personality. Features from the MRC psycholinguistic database produce the best ranking model for emotional stability, whereas LIWC features are the most useful for openness to experience. Finally, ranking models for agreeableness and conscientiousness perform best when combining all feature sets together.

Should personality be modelled as a set of continuous or discrete dimensions?

Although we successively treat the personality recognition task as a classification, regression and ranking problem, results cannot be compared directly as different evaluation metrics are used. Nevertheless, Section 3.6 explores this question by assessing the performance of ranking models in a binary classification task, as well as the performance of multi-class classification models in a ranking task. Results in Table 3.25 show that the ranking models significantly outperform the best performing classifier in the ranking task for four traits out of five, while the ranking models do not perform significantly worse than the best classifier in the classification task.[1] Although additional experiments would be required to reach significance, these exploratory results suggest that continuous ranking models provide a better fit to our data.

Personality generation:

How can personality markers identified in psychological studies be reproduced in a natural language generation system?

While previous work explores the use of n-gram models to rank output utterances [Isard et al., 2006], our framework for controlling the system's personality relies on the PERSONAGE base generator, which implements generation decisions that are derived from psychology findings about how personality traits affect human language production (see Chapters 4 and 5). We investigate different techniques for controlling the personality produced by PERSONAGE. Chapter 6 evaluates PERSONAGE-RB, a rule-based version of our generator in which extreme personality is generated using predefined parameter settings suggested by psychological studies. We also evaluate two data-driven methods for generating personality varying along a continuous scale. Chapter 8 presents results obtained by using the overgenerate and select paradigm—implemented in the PERSONAGE-OS stochastic generator—by learning statistical models that select one or more utterances matching the target personality out of a candidate utterance set. Finally, Chapter 9 presents a novel approach to data-driven generation that does not overgenerate, by learning parameter estimation models that predict the generation parameters to convey any combination of

[1] The comparison with a traditional classification task is not strictly correct. See page 72.

personality scores. This technique is implemented in the PERSONAGE-PE generator.

Can these personality markers be used to convey recognisable personality in a specific application domain?

We evaluate personality generation methods by assessing whether they generate recognisable personality variation, according to human judges. Results in Table 6.3 in Chapter 6 show that judges can discriminate between the personality traits conveyed by PERSONAGE-RB's rule-based utterances, for each of the Big Five dimensions. The best results are obtained for extraversion, as extravert utterances are rated on average 3.02 points higher than introvert utterances, on a scale from 1 to 7. Openness to experience is the hardest trait to project, with an average rating difference of 1.32 between the two utterance sets. These significant results suggest that PERSONAGE-RB could be used in applications requiring a small number of extreme personality traits. We developed and tested this in a sample 'concierge' application, where we focused on the linguistic output requirements for a computational system to make recommendations and comparisons of restaurants, with the character able to manifest extreme personality traits when presenting such information. In Section 10.2, we present an example showing that some of our techniques can be used in other types of dialogue applications, such as in interactive drama systems to modify or 'improvise' on the character's original dialogue.

Can the perception of multiple personality traits be controlled continuously using statistical models trained on personality judgements of generated utterances?

We implement and evaluate two data-driven generation techniques, in order to generate personality varying on a continuous scale. First, the overgenerate and select approach is implemented in the PERSONAGE-OS data-driven generator. While previous work using this method was not evaluated [Isard et al., 2006], Table 8.1 in Chapter 8 shows that selection models trained on randomly generated utterances predict the personality of unseen utterances with correlations ranging from .19 (openness to experience) to .54 (agreeableness), significantly outperforming the random selection baseline. When including psychological knowledge—by adding utterances generated using parameter settings derived from psychology findings

(20% of the dataset)—correlations increase to a minimum of .31 for openness to experience, and to a maximum of .66 for emotional stability. These results suggest the need for hybrid approaches combining existing knowledge and stochastic exploration.

The PERSONAGE-PE generator implements a second data-driven method presented in Chapter 9, using parameter estimation models to control the generation process. A large-scale evaluation using naive judges shows in Table 9.6 that the average ratings correlate strongly with the target personality scores for all traits but conscientiousness, with correlation coefficients ranging from .41 (openness to experience) up to .80 (extraversion). Finally, results show that judges successfully perceive fine-grained variation (within a unit on a scale from 1 to 7) for both ends of the emotional stability scale, and for the positive end of the openness to experience scale.

Can general-domain personality recognition models successfully predict the personality of generated utterances?

As part of the evaluation of the overgenerate and select method, Chapter 8 reports experiments using the personality recognition models trained on general-domain data in Chapter 3 for selecting the output utterance [Mairesse et al., 2007]. Results summarised in Table 8.6 show that models trained on observer reports of conversations predict the extraversion of randomly generated utterances better than the baseline ($r = .25$), while models trained on self-reports of essays are the best predictors of agreeableness ($r = .33$) and conscientiousness ($r = .20$). While out-of-domain models are less accurate than models trained on PERSONAGE's outputs, they can be re-used in different application domains. Moreover, Table 8.6 shows that the accuracy of out-of-domain models increases when adding utterances generated using PERSONAGE-RB to the candidate utterance set, suggesting that these models would be useful in a hybrid approach that provides a mechanism for including the rule-based knowledge derived from psychology findings into the overgeneration phase of the overgenerate and select approach.

Can the system project a target personality through a single utterance, or is more information needed?

In Chapter 4, we hypothesised that personality could be recognised by manipulating the linguistic cues of a single utterance. Results presented in Chapter 6 confirm the validity of this assumption, as the judges recognise the utterance's intended personality for each end of each Big Five scale. However, it is likely that additional cues—e.g. characterising the system's dialogue strategy—would improve our results, as long as they are used in a consistent way.

What generation decisions are the most useful for conveying a specific personality trait?

Chapter 7 evaluates how generation decisions affect the perception of the utterance's personality. This analysis confirms many hypotheses made in Chapter 4, such as the association of extraversion with verbosity. Some hypothesised markers do not generalise, e.g. whereas Pennebaker and King [1999] find that extraverts use more positive emotion words, positive content is not associated with extraversion in our domain. This analysis also reveals personality markers that were not previously identified in the literature, e.g. disagreeableness is marked by the use of concessions, and in-group markers indicate openness to experience. The interested reader is referred to Section 7.2.1 for more detailed results.

How do data-driven generation techniques compare to a rule-based approach purely based on findings from the psychology literature?

Concerning the overgenerate and select method, Chapter 7 shows in Section 7.1.1 that the random overgeneration produces utterances that are perceived as more extreme than those generated using PERSONAGE-RB's predefined parameter settings. However, Section 8.3.1 shows that extreme personality is modelled less accurately than mid-range personality, because of the few number of extreme utterances in the selection models' training data. Furthermore, Section 8.3.2 shows that the generation of extreme utterances requires a relatively large candidate set, as well as selection models that produce a large variation in their output.

Parameter estimation models are compared to the rule-based method in Section 9.2.3, showing that they are not generally able to produce utterances perceived as extreme. Thus, we find that data-driven methods do not perform better than a psychologically-informed rule-based generator for the generation of extreme personality. We believe this is due to the normal distribution of personality ratings in the training data, resulting in few extreme utterances to learn from. Nevertheless, data-driven methods offer a principled approach for generating continuous variation over all Big Five dimensions simultaneously.

10.2 Generalisation to other domains

This thesis shows that PERSONAGE can convey personality in the restaurant recommendation domain. We believe our work can be trivially extended to any tourist domain (e.g. hotels), and more generally to any domain producing evaluative utterances (e.g. film reviews). As an extension to these domains would simply be a matter of keyword substitution in PERSONAGE's output, we would like to assess whether PERSONAGE is likely to generalise to more general domains, involving any kind of content and dialogue act. In this section, we thus explore PERSONAGE's generalisation capability in an interactive narrative system (INS).[2]

An INS consists of a textual or graphical environment in which a human participant interacts with fictional characters, typically using unconstrained language. The participant's input affects the storyline in real-time, by triggering specific responses from the characters based on their emotional state and personality. Thus, INS have the technical requirement of producing outputs that are recognisable as manifesting differences in dramatic character. While natural language processing techniques are typically used to understand a wide range of user inputs, the utterances produced by the system are often highly handcrafted, e.g. pre-recorded by actors. Although this technique produces high quality utterances, it suffers from the following weaknesses:

- Recording all possible system outputs is expensive and time-consuming;
- It does not scale well to new domains;
- For each dialogue act, a new prompt must be recorded for every possible

[2]INS are also referred to as interactive story-telling systems.

pragmatic variation in the system, e.g. depending on the utterance's context and the speaker's state.

These issues are referred to as the *authoring bottleneck* [Mateas, 2007], which could be reduced by moving towards automated methods. The automation of various aspects of INS could ultimately lead to off-the-shelf tools being available, allowing the creation of interesting characters without dealing with implementation details. We therefore suggest that the PERSONAGE generator presented in this thesis is a first step in the direction of a general framework for automatically generating character variation in a comprehensive way.

To explore this claim, we use the PERSONAGE-RB generator to manipulate the personality conveyed by the characters of FAÇADE, a three-dimensional INS in which the player is invited for dinner by Trip and Grace, a dysfunctional couple who quickly expose their marriage issues [Mateas and Stern, 2003].

An important choice faced by INS developers is the level of abstraction of the manipulated language. FAÇADE currently produces pre-recorded prompts, i.e. the variation is hand-coded in each utterance. We investigate deeper representations by manually converting dialogue turns into syntactic structures and modifying them using PERSONAGE-RB's pragmatic marker insertion and lexical choice components (see Chapter 5). Tables 10.1 and 10.2 show a series of modified dialogues between Trip and Grace. Each column represents the same initial dialogue, in which both speakers were assigned a different personality.

Although there is no generation at the content planning and aggregation levels, the amount of variation is still substantial, see Trip changing from an extravert in the third column of Table 10.1, i.e. *'Yeah, Grace's decorating has been such like, a really great outlet for her, you know!'*, to a more serious conscientious person in the third column of Table 10.2, i.e. *'I see, ok, I suppose Grace's decorating has been such a rather outstanding outlet for her'*. On the other end of the same dimensions, Grace produces an introvert answer in Table 10.1, i.e. *'Outlet? Err... mmhm... what do you mean by an outlet?'*, whereas her unconscientious answer in Table 10.2 reflects hesitancy and carelessness, i.e. *'I don't know mate! Outlet? Mmhm... err... what do you mean by an outlet?'*. While the perception of the personality in this new domain remains to be evaluated, we believe these example dialogues provide an argument for automating variation in INS, by manipulating deep utterance representations.

Speaker	Original dialogue	TRIP: extravert GRACE: introvert	TRIP: emotionally stable GRACE: neurotic
USER	· Do you like this place?	· Do you like this place?	· Do you like this place?
TRIP	· Yeah, you know, Grace's decorating has been such a great outlet for her.	· Yeah, Grace's decorating has been such like, a really great outlet for her, you know!	· Basically, Grace's decorating has been such a kind of outstanding outlet for her.
GRACE	· Outlet? What do you mean outlet?	· Outlet? Err... mmhm... what do you mean by an outlet?	· I mean, actually, I am not really sure. Outlet? Err... what do you mean by an ou-outlet?
TRIP	· And I am always encouraging her to buy. Only the best for us, right Grace?	· Actually, I am always encouraging her to buy buddy, you know! Only the best for us.	· Basically, I am always encouraging her to buy. Only the best for us.
GRACE	· Trip, let's not get too carried away here.	· Well, right, let's not get too quite carried away here.	· Actually, I am not re-really sure! Err... I mean, let's not get too carried away here.
TRIP	· Aha! Great! I will fix us some drinks. Good news, I just fully loaded the bar, with the real quality.	· Oh gosh actually, I will fix us some drinks mate, wouldn't you? I fully just loaded the bar with really kind of real quality.	· I will fix us some drinks. Basically, I fully just loaded the bar with sort of real quality.
GRACE	· Trip, please... let's not go overboard with the drink preparation. The funny thing is, Trip doesn't even like the taste of alcohol.	· Well, i see, let's not go sort of overboard with the drink preparation. I guess the rather funny thing is, Trip doesn't even like the taste of alcohol.	· I don't know! Ac-ac-actually, let's not go really overboard with the drink preparation. I mean, the funny thing is err... Trip doesn't even like the taste of alcohol.
TRIP	· How about you Grace, can I pour you a glass? I assume you want your usual... 'a lovely, very cold glass of Chardonnay'.	· How about you Grace. Can I pour you a glass? I assume you want like, a really lovely very cold glass of Chardonnay, pal, you know!	· How about you Grace. Can I pour you a glass? Basically, I assume you want a quite lovely very sort of cold glass of Chardonnay.
GRACE	· Um, I'd like a simple glass of Chardonnay, thank you.	· Err... right, I would like a sort of simple glass of Chardonnay.	· I don't know! Err... I mean, I would like like, a really si-simple glass of Chardonnay.

Table 10.1: Example of PERSONAGE-RB's outputs for extraversion and emotional stability, based on a script from FAÇADE [Mateas and Stern, 2003]. The original dialogue was manually converted into a sequence of syntactic structures, and processed by PERSONAGE-RB's pragmatic marker insertion and lexical choice components (see Chapter 5).

Speaker	**TRIP: agreeable GRACE: disagreeable**	**TRIP: conscientious GRACE: unconscient.**	**TRIP: open GRACE: not open**
USER	· Do you like this place?	· Do you like this place?	· Do you like this place?
TRIP	· Well, yeah, Grace's decorating has been such a sort of outstanding outlet for her mate, you know, alright?	· I see, ok, I suppose Grace's decorating has been such a rather outstanding outlet for her.	· I see, I suppose Grace's decorating has been such a sort of outstanding outlet for her, you know, you see?
GRACE	· Outlet? What I mean, do you mean by an outlet?	· I don't know mate! Outlet? Mmhm... err... what do you mean by an outlet?	· I mean, basically, I am not sure. Outlet? Err... mmhm... what do you mean by an outlet?
TRIP	· Ok, right, I am always encouraging her to buy, you know mate, okay? Only the best for us.	· Well, i see, I am always encouraging her to buy. Only the best for us.	· I see, mmhm... I am always encouraging her to buy, you know, okay? Only the best for us.
GRACE	· Oh God I mean, let's not get too quite carried away here.	· Yeah, I mean, let's not get too rather carried away here mate!	· Err... mmhm... I am not sure. I mean, let's not get too carried away here.
TRIP	· Ok, oh I will fix us some drinks buddy, wouldn't you? I fully just loaded the bar with rather real quality, you know.	· I see, I will fix us some drinks. I fully just loaded the bar with rather real quality.	· I will fix us some drinks, you know, you see? I fully just loaded the bar with quite real quality.
GRACE	· Actually, basically, let's not go overboard with the drink preparation. I mean, the funny thing is mmhm... Trip doesn't even like the taste of alcohol.	· I mean, I am not sure. Mmhm... let's not go overboard with the drink preparation! The kind of funny thing is, Trip doesn't even like the taste of alcohol, pal.	· I am not sure. Err... I mean, let's not go overboard with the drink preparation. The funny thing is mmhm... Trip doesn't even like the taste of alcohol.
TRIP	· How about you Grace. Can I pour you a glass? I assume you want a sort of lovely very quite cold glass of Chardonnay buddy, you know, alright?	· How about you Grace. Can I pour you a glass? I assume you want a sort of lovely very quite cold glass of Chardonnay.	· How about you Grace. Can I pour you a glass? I assume you want a quite lovely very sort of cold glass of Chardonnay, you know, okay?
GRACE	· Actually, I mean, I would like a kind of simple glass of Chardonnay.	· I mean, I am not sure. Err... I would like a kind of simple glass of Chardonnay, you know buddy.	· Mmhm... I mean, I don't know. Err... I would like like, a simple glass of Chardonnay.

Table 10.2: Example outputs for agreeableness, conscientiousness and openness to experience, from the script from FAÇADE in Table 10.1 [Mateas and Stern, 2003].

As FAÇADE's original characters already have specific personalities (see leftmost dialogue in Table 10.1), a first step was to remove any marker of personality to produce neutral syntactic structures. Personality is also marked at the content level, e.g. *'Trip, let's not get too carried away here'*, which can lead to unnatural utterances when inconsistent markers are introduced. This suggests that personality needs to be controlled in a consistent way at all levels of language production as we do in the information presentation domain, if it is to produce realistic outputs.

Nevertheless, this section shows that manipulations at the syntactic level can produce interesting variation. We believe that this variation could be made more consistent using a deeper input representation (e.g. a content plan), which could be provided by authors during the INS design phase.

While this section presents exploratory work, we hope to give the reader a sense of the possibilities offered by stylistic language generation.

10.3 Future research

As mentioned in Chapter 1, this research focuses on methods for personality recognition and generation intended for dialogue applications such as interactive drama systems, spoken dialogue systems and intelligent tutoring systems. While the evaluation of these methods is an important first step towards automatic adaptation in dialogue, some of the challenges discussed in the introduction remain. An important next step is to use these techniques to simultaneously model the personality of the user and the system in dialogue, in order to test various hypotheses regarding personality-based alignment, such as the similarity-attraction effect suggested by Reeves and Nass [1996]. Furthermore, the optimal personality of the system is likely to be application-dependent (see Section 1.4.2), it would thus be useful to evaluate how the user's and the system's personality affect task performance in different applications.

Future work on personality recognition models should further evaluate their generalisation to new domains. Results in Chapter 8 suggest that they do generalise to the information presentation domain, although they perform worse than models trained on in-domain data. It is also not clear whether the accuracies are high enough to be useful. Applications involving speech recognition will introduce noise in all features except for the prosodic features, probably reducing model accuracy, but we expect that more training data would improve performance. While our experiments suggest that models trained on stream-of-consciousness essays do not have a large predictive power, it is not clear whether this is due to the nature of written language, the self-reports of personality, or the combination of both. Further data would thus be needed to fully evaluate the impact of the language source on modelling accuracy, by collecting observer ratings for the stream-of-consciousness essays for example. Additionally, we believe that the inclusion of gender as a feature would produce better models, as language correlates of perceived personality were shown to depend on the gender of the speaker [Mehl et al., 2006]. Another issue is the poor performance of the utterance type features, possibly due to our rudimentary automated labelling technique (see Section 3.2.2). As we believe that important personality markers can be found at the speech act level, future work should include more advanced speech act detection techniques as part of the feature extraction phase. Finally, while Chapter 3 focuses purely on data-driven methods, it would be interesting to evaluate whether the addition of knowledge derived from psychology findings improves recognition accuracy, as done in Chapter 4 for the personality generation task.

Concerning the generation of the system's personality, our approach could be extended to other aspects of language production, such as dialogue strategy selection and prosody. As findings suggest that personality affects both aspects of dialogue [Vogel and Vogel, 1986, Scherer, 1979, *inter alia*], our methodology could be applied to the parameterisation of a dialogue manager and a text-to-speech engine, in order to project a consistent personality to the user.

Future work should also address why Chapter 6's rule-based method generally performs better for projecting extreme personality, while producing more natural utterances. Although PERSONAGE-RB only generates extreme personality traits, interpolation techniques could be used to generate finer-grained personality variation. While parameter estimation models provide an efficient and principled data-driven technique for generating continuous linguistic variation, further experiments should evaluate why they fail to project conscientiousness correctly, e.g. by using a larger sample of judges.

Some of the psychologically-motivated generation decisions do not correlate enough with the judges' ratings to be included in the parameter estimation models (see Section 9.1.4 in Chapter 9). This could be due to the relatively small development dataset size (160 utterances to optimise 67 parameters), or to the implementation of some parameters. As the parameter-independence assumption could also be responsible, statistical models that simultaneously learn multiple dependent variables might improve performance. However, increasing the size of the output space could further aggravate data sparsity issues.

Although most of PERSONAGE's variation is generated automatically, the system assumes the existence of a generation dictionary containing syntactic templates that express various pragmatic effects (e.g. different connotations or polarity, see Section 5.5.2 in Chapter 5). While our dictionary is currently handcrafted, other research has started to investigate methods for extracting the generation dictionary from data [Higashinaka et al., 2007].

We have shown that evaluative utterances in the restaurant domain can manifest personality, but more research is needed to identify which speech acts recognisably manifest personality in a restricted domain. Although PERSONAGE's parameters were implemented with domain-independence in mind, future work should assess the extent to which the parameters derived from psychological findings are dependent on these speech acts, as well as on the application domain. The application to interactive narrative systems presented in the last section suggests potential benefits of domain-independent generation of personality for the entertainment industry.

Finally, the techniques presented in this thesis model personality as a descriptive mapping between linguistic markers and personality scores. Further studies of the causes of personality traits could lead to the development of *generative* models reproducing mechanisms identified in human beings (see Section 2.1.2), which could be used to predict unseen personality markers.

10.4 Conclusion

This thesis has presented and evaluated various techniques for modelling individual differences in language production. We believe that dialogue applications such as spoken dialogue systems, interactive drama systems and intelligent tutoring systems would benefit from taking these differences into account, and that personality traits represent an appropriate set of dimensions for mediating this adaptation. While data-driven techniques offer a principled method for learning these individual differences, we find that psychological knowledge based on more general data is necessary for (1) implementing generation parameters that affect the utterance's personality (see Chapter 4), (2) informing the generation process for producing recognisable *extreme* personality (see Chapter 6), and (3) using these extreme utterances for improving the coverage of stochastic data-driven methods (see Chapter 8).

We provide a fully implemented and evaluated example of a personality generation capability for dialogue applications that is completely generative. Starting from a meaning representation, we show how personality affects all phases of the language generation process, and that certain parameters such as the polarity of the content selected have a strong effect on the perception of personality.

Appendix A

Utterances Generated using PERSONAGE-RB

This appendix contains the utterances generated using PERSONAGE-RB for the evaluation presented in Chapter 6. It contains 40 utterances for each end of the extraversion scale, and 20 utterances for each end of the emotional stability, agreeableness, conscientiousness and openness to experience dimensions. There are thus 240 utterances in total.

Content plan	Trait	End
compare(Aureole, Daniel)	**extraversion**	**low**
Daniel is costly. Aureole features beautiful ambience. I think that Daniel and Aureole are the only restaurants that are any good.		
compare(Aureole, Daniel)	**extraversion**	**low**
Right, Daniel's price is 82 dollars. Aureole, however, features marvelous waiters. I might recommend Daniel and Aureole.		
compare(Da Andrea, John's Pizzeria, Uguale)	**extraversion**	**low**
Da Andrea is costly while John's Pizzeria doesn't feature nasty waiters. John's Pizzeria, Da Andrea and Uguale are satisfactory restaurants.		
compare(Da Andrea, John's Pizzeria, Uguale)	**extraversion**	**low**
I mean, Uguale isn't cheap. On the other hand, Da Andrea features acceptable ambience. John's Pizzeria, Da Andrea and Uguale are satisfactory restaurants.		
compare(Daily Soup, John's Pizzeria)	**extraversion**	**low**
Daily Soup features mediocre ambience and rude staff. However, John's Pizzeria doesn't have unfriendly waiters. Err... I mean, John's Pizzeria and Daily Soup are the only restaurants that are any good.		
compare(Daily Soup, John's Pizzeria)	**extraversion**	**low**
Daily Soup doesn't feature good ambiance. However, John's Pizzeria has welcoming waiters. John's Pizzeria and Daily Soup are the only eating places that are any good.		
compare(Dishes, Joe's Shanghai)	**extraversion**	**low**
Dishes doesn't have good atmosphere. Joe's Shanghai, on the other hand, is low-cost. Joe's Shanghai and Dishes are the only places that are any good.		
compare(Dishes, Joe's Shanghai)	**extraversion**	**low**
Right, I think that Dishes doesn't have good ambience. Joe's Shanghai, which features mediocre ambiance, is a chinese eating place. Its price is 23 dollars. There could be worse restaurants.		
compare(Dojo, Japonica)	**extraversion**	**low**
Well, I think that Dojo has mediocre ambiance. On the other hand, Japonica doesn't feature nasty ambience.		
compare(Dojo, Japonica)	**extraversion**	**low**
It seems that Dojo doesn't have good ambiance. On the other hand, Japonica features satisfactory ambience.		
compare(Dojo, Japonica)	**extraversion**	**low**
Well, I mean, Japonica's price is 37 dollars. However, Dojo doesn't feature nasty food.		
compare(Dojo, Japonica)	**extraversion**	**low**
Dojo doesn't feature nasty food. I mean, Japonica has acceptable ambience.		
compare(Dojo, Japonica)	**extraversion**	**low**
Dojo doesn't have good ambiance. It seems that Japonica, on the other hand, features acceptable ambience.		
compare(Dojo, Japonica)	**extraversion**	**low**
Japonica is a japanese and sushi restaurant while Dojo doesn't feature nasty food.		
compare(Ferrara, Joe's Shanghai, Mangia)	**extraversion**	**low**
Joe's Shanghai features poor ambiance. On the other hand, Ferrara doesn't have nasty ambience. Ferrara, Mangia and Joe's Shanghai are the only restaurants that are any good.		
compare(Ferrara, Joe's Shanghai, Mangia)	**extraversion**	**low**
Although Joe's Shanghai's price is 23 dollars, it features poor ambiance while Mangia has acceptable ambience. Ferrara, Mangia and Joe's Shanghai are the only restaurants that are any good.		
compare(Lemongrass Grill, Monsoon)	**extraversion**	**low**
Well, right, Lemongrass Grill, which doesn't feature good ambience, is a thai restaurant. Monsoon, on the other hand, has acceptable ambiance.		
compare(Lemongrass Grill, Monsoon)	**extraversion**	**low**
It seems to me that Lemongrass Grill has mediocre ambiance. Monsoon doesn't, however, feature nasty ambience.		
compare(Lemongrass Grill, Monsoon)	**extraversion**	**low**
I think that Lemongrass Grill features mediocre ambience. Monsoon doesn't, on the other hand, have nasty ambiance.		
compare(Lemongrass Grill, Monsoon)	**extraversion**	**low**
I think that Lemongrass Grill has poor ambiance. On the other hand, Monsoon doesn't feature nasty ambience.		
recommend(Amy's Bread)	**extraversion**	**low**
Amy's Bread is the only restaurant that is any good.		
recommend(Amy's Bread)	**extraversion**	**low**
Amy's Bread's price is 12 dollars. It isn't as bad as the others.		
recommend(Amy's Bread)	**extraversion**	**low**
Amy's Bread, which is the only restaurant that is any good, isn't expensive.		
recommend(Amy's Bread)	**extraversion**	**low**
I mean, Amy's Bread isn't as bad as the others.		

<table>
<tr><th>Content plan</th><th>Trait</th><th>End</th></tr>
<tr><td>recommend(Azuri Cafe)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Azuri Cafe’s price is 14 dollars. It’s the only restaurant that is any good.</td></tr>
<tr><td>recommend(Azuri Cafe)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Azuri Cafe is the only place that is any good.</td></tr>
<tr><td>recommend(Bond Street)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Bond Street is the only restaurant that is any good.</td></tr>
<tr><td>recommend(Bond Street)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">It seems that Bond Street is the only restaurant with friendly staff that is any good.</td></tr>
<tr><td>recommend(Chanpen Thai)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">It seems to me that Chanpen Thai is the only restaurant with acceptable ambience that is any good.</td></tr>
<tr><td>recommend(Chanpen Thai)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Right, Chanpen Thai is the only eating house that is any good.</td></tr>
<tr><td>recommend(Flor De Mayo)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Flor De Mayo isn’t as bad as the others.</td></tr>
<tr><td>recommend(Flor De Mayo)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Flor De Mayo isn’t as bad as the others with adequate staff.</td></tr>
<tr><td>recommend(John’s Pizzeria)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Well, I mean, John’s Pizzeria is the only restaurant that is any good.</td></tr>
<tr><td>recommend(John’s Pizzeria)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">I think that John’s Pizzeria isn’t as bad as the others.</td></tr>
<tr><td>recommend(Le Marais)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Right, I mean, Le Marais is the only restaurant that is any good.</td></tr>
<tr><td>recommend(Le Marais)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Err... it seems to me that Le Marais isn’t as bad as the others.</td></tr>
<tr><td>recommend(River)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">I think that River is the only restaurant that is any good.</td></tr>
<tr><td>recommend(River)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">I mean, River isn’t err... as bad as the others, you know.</td></tr>
<tr><td>recommend(Ruby Foo’s)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">I mean, it seems that Ruby Foo’s isn’t as bad as the others.</td></tr>
<tr><td>recommend(Ruby Foo’s)</td><td>extraversion</td><td>low</td></tr>
<tr><td colspan="3">Ruby Foo’s is the only restaurant that is any good.</td></tr>
<tr><td>compare(Aureole, Daniel)</td><td>extraversion</td><td>high</td></tr>
<tr><td colspan="3">I am sure you would like Daniel and Aureole. Daniel just has wonderful servers and the ambience is lovely. The food is kind of brilliant, even if it’s expensive. Aureole features great service and the atmosphere is beautiful. The food is excellent, even if it’s costly.</td></tr>
<tr><td>compare(Aureole, Daniel)</td><td>extraversion</td><td>high</td></tr>
<tr><td colspan="3">I am sure you would like Daniel and Aureole, you know. Aureole has really great food, the servers are wonderful and the atmosphere is beautiful. Daniel has great service with lovely atmosphere. Even if it’s expensive, the food is excellent.</td></tr>
<tr><td>compare(Da Andrea, John’s Pizzeria, Uguale)</td><td>extraversion</td><td>high</td></tr>
<tr><td colspan="3">I would recommend John’s Pizzeria, Da Andrea and Uguale, wouldn’t you? Da Andrea has great servers and the atmosphere is acceptable. Uguale has nice atmosphere. Actually, the service is good, even if its price is 33 dollars. Even if John’s Pizzeria just has poor atmosphere, its price is around 20 dollars. The servers are friendly.</td></tr>
<tr><td>compare(Da Andrea, John’s Pizzeria, Uguale)</td><td>extraversion</td><td>high</td></tr>
<tr><td colspan="3">I am sure you would like John’s Pizzeria, Da Andrea and Uguale. Basically, Da Andrea features good staff, the servers are great, and the ambiance is acceptable. Uguale has nice ambience. Even if its price is 33 dollars, the service is friendly. Even if John’s Pizzeria has poor atmosphere, its price is 20 dollars. Actually, the service is adequate.</td></tr>
<tr><td>compare(Daily Soup, John’s Pizzeria)</td><td>extraversion</td><td>high</td></tr>
<tr><td colspan="3">I am sure you would like John’s Pizzeria and Daily Soup. John’s Pizzeria’s price is around 20 dollars and the service is friendly. It has like, good food, even if the atmosphere is poor. Daily Soup’s price is 10 dollars, even if the ambience is bad, you know. The food is acceptable and the staff is rude.</td></tr>
<tr><td>compare(Daily Soup, John’s Pizzeria)</td><td>extraversion</td><td>high</td></tr>
<tr><td colspan="3">I am sure you would like John’s Pizzeria and Daily Soup. John’s Pizzeria is inexpensive with friendly waiters, isn’t it? The food is good, even if the ambience is bad. Daily Soup is cheap, even if it has poor atmosphere. Even if the servers are rude, basically, the food is just nice.</td></tr>
</table>

Content plan	Trait	End
compare(Dishes, Joe's Shanghai)	**extraversion**	**high**
I would recommend Joe's Shanghai and Dishes. Dishes is cheap, even if the atmosphere is poor. It's a new american place with good food. Actually, Joe's Shanghai's price is 23 dollars and it's a chinese restaurant. The food is nice. Basically, the ambience is just bad.		
compare(Dishes, Joe's Shanghai)	**extraversion**	**high**
Yeah, I would recommend Joe's Shanghai and Dishes, wouldn't you? Basically, Dishes has good food. Actually, its price is 14 dollars, even if the atmosphere is poor. Joe's Shanghai is affordable and it's a kind of chinese place. Even if the atmosphere is bad, it has good food, good food.		
compare(Dojo, Japonica)	**extraversion**	**high**
Actually, Japonica has friendly waiters and acceptable ambience, pleasant atmosphere, it's a japanese and sushi place, you know, Dojo is a japanese and vegetarian restaurant, the servers are rude and the ambiance is kind of poor.		
compare(Dojo, Japonica)	**extraversion**	**high**
Japonica just has kind of nice servers, the atmosphere is acceptable, Dojo is a japanese and vegetarian place, the waiters are bad and the ambience is poor.		
compare(Dojo, Japonica)	**extraversion**	**high**
Dojo's price is around 14 dollars and it has poor atmosphere. Even if Japonica is expensive, the food is good, you know. Basically, the ambience is nice.		
compare(Dojo, Japonica)	**extraversion**	**high**
Japonica has kind of good food, even if its price is 37 dollars. Even if Dojo just has poor atmosphere, its price is 14 dollars. The food is nice, the food is just adequate.		
compare(Dojo, Japonica)	**extraversion**	**high**
Yeah, basically, Japonica features kind of good food, the atmosphere is acceptable and it's a japanese and sushi place. Dojo has adequate food, even if the ambience is really poor.		
compare(Dojo, Japonica)	**extraversion**	**high**
Yeah, Japonica has nice food, the atmosphere is pleasant and it's a japanese and sushi place. Dojo is a japanese and vegetarian restaurant, isn't it? Even if the ambience is poor, actually, the food is kind of acceptable.		
compare(Ferrara, Joe's Shanghai, Mangia)	**extraversion**	**high**
I am sure you would like Ferrara, Mangia and Joe's Shanghai. Ferrara's price is 17 dollars, you know, the food is good, Mangia just has good food, its price is around 20 dollars, Joe's Shanghai has good food and it's low-cost.		
compare(Ferrara, Joe's Shanghai, Mangia)	**extraversion**	**high**
Basically, I am sure you would like Ferrara, Mangia and Joe's Shanghai. Ferrara is cheap, its price is around 17 dollars, it just has dainty food and nice ambience, Mangia features tasty food, it's inexpensive, the ambiance is acceptable, Joe's Shanghai's price is 23 dollars and it has good food and poor atmosphere.		
compare(Lemongrass Grill, Monsoon)	**extraversion**	**high**
Actually, Monsoon is a vietnamese restaurant with nice ambience, Lemongrass Grill is a thai place, you know and it has poor atmosphere.		
compare(Lemongrass Grill, Monsoon)	**extraversion**	**high**
Basically, Monsoon has acceptable atmosphere, the atmosphere is nice, you know, it's a vietnamese place, Lemongrass Grill is a thai restaurant and the ambience is kind of poor.		
compare(Lemongrass Grill, Monsoon)	**extraversion**	**high**
Lemongrass Grill's price is 22 dollars, even if the ambiance is poor. Monsoon is kind of low-cost and the atmosphere is acceptable, it features pleasant ambience.		
compare(Lemongrass Grill, Monsoon)	**extraversion**	**high**
Yeah, Lemongrass Grill's price is 22 dollars, even if the ambience is just really poor. Monsoon is low-cost and the atmosphere is nice.		
recommend(Amy's Bread)	**extraversion**	**high**
I am sure you would like Amy's Bread. Basically, its price is 12 dollars, it's cheap, you know, the food is good and the servers are friendly.		
recommend(Amy's Bread)	**extraversion**	**high**
I am sure you would like Amy's Bread. Even if the atmosphere is poor, it's cheap. It has good food and kind of nice servers, you know.		
recommend(Amy's Bread)	**extraversion**	**high**
Basically, I am sure you would like Amy's Bread, you know. Its price is 12 dollars, the food is really kind of good, the servers are nice and it's a cafes place.		
recommend(Amy's Bread)	**extraversion**	**high**
I am sure you would like Amy's Bread, it's one of my favourite places, you know. Its price is around 12 dollars. The food is just good. It's in Midtown West and a cafes restaurant with nice servers.		

Content plan	Trait	End
recommend(Azuri Cafe)	**extraversion**	**high**
Basically, I am sure you would like Azuri Cafe, it's one of my favourite places. Even if the atmosphere is bad, it's cheap. Actually, it just has good food and bad servers.		
recommend(Azuri Cafe)	**extraversion**	**high**
I am sure you would like Azuri Cafe. The food is just good, it's a kind of kosher and vegetarian place and the servers are bad. It's cheap, even if it has poor atmosphere.		
recommend(Bond Street)	**extraversion**	**high**
I am sure you would like Bond Street, you know. Basically, the food is great and the atmosphere is good with friendly service.		
recommend(Bond Street)	**extraversion**	**high**
Yeah, Bond Street is the best place. The atmosphere is good, it has nice service and it's a japanese and sushi place. Even if it's expensive, you know, the food is great.		
recommend(Chanpen Thai)	**extraversion**	**high**
Actually, I am sure you would like Chanpen Thai, it's the best place. The service is nice, the food is just kind of adequate, its price is 24 dollars, it has acceptable atmosphere and it's a thai restaurant.		
recommend(Chanpen Thai)	**extraversion**	**high**
I am sure you would like Chanpen Thai, you know. Actually, the waiters are friendly, the staff is acceptable, and it's low-cost with really adequate food and pleasant ambience.		
recommend(Flor De Mayo)	**extraversion**	**high**
I am sure you would like Flor De Mayo, you know. The food is kind of good, the food is tasty, it has nice servers, it's in Uptown Manhattan and it's a chinese and latin american place. Its price is around 18 dollars, even if the atmosphere is poor.		
recommend(Flor De Mayo)	**extraversion**	**high**
Flor De Mayo is one of my favourite restaurants, isn't it? It just has really good food, the service is nice and it's just located in Uptown Manhattan and a chinese and latin american place. Even if the atmosphere is poor, its price is 18 dollars.		
recommend(John's Pizzeria)	**extraversion**	**high**
I am sure you would like John's Pizzeria, it's one of my favourite places. It's cheap. Even if the atmosphere is just bad, it has really good food.		
recommend(John's Pizzeria)	**extraversion**	**high**
I am really sure you would like John's Pizzeria. Its price is 20 dollars. Even if the atmosphere is bad, actually, it has good food, good food.		
recommend(Le Marais)	**extraversion**	**high**
Basically, actually, I am sure you would like Le Marais. It features friendly service and acceptable atmosphere and it's a french, kosher and steak house place. Even if its price is 44 dollars, it just has really good food, nice food.		
recommend(Le Marais)	**extraversion**	**high**
I am sure you would like Le Marais, you know. The atmosphere is acceptable, the servers are nice and it's a french, kosher and steak house place. Actually, the food is good, even if its price is 44 dollars.		
recommend(River)	**extraversion**	**high**
Actually, I am sure you would like River. The food is acceptable, it has nice service, the atmosphere is pleasant and it's kind of expensive.		
recommend(River)	**extraversion**	**high**
I am really sure you would like River. Actually, the food is nice, the service is friendly, it's a thai and vietnamese place with acceptable atmosphere and it isn't cheap.		
recommend(Ruby Foo's)	**extraversion**	**high**
I am really sure you would like Ruby Foo's, you know. The atmosphere is just good with tasty food and it's a chinese, japanese and thai place with nice service.		
recommend(Ruby Foo's)	**extraversion**	**high**
I am sure you would like Ruby Foo's, it's one of my favourite places. Actually, the atmosphere is just good, it has good food and the service is nice.		
compare(Acacia, Marinella)	**emotional stability**	**low**
I might be wrong. I might approve Acacia and Marinella. Err... Acacia provides like, acceptable food. It's in Midtown! It's a bistro and french place. Actually, I mean, Marinella is in Manhattan and an it-it-italian place.		
compare(Caffe Cielo, Trattoria Spaghetto)	**emotional stability**	**low**
I am not sure! Actually, I might recommend Caffe Cielo and Trattoria Spaghetto. Caffe Cielo, which has poor atmosphere, bad atmosphere, even if it features like, nice food, is an italian re-restaurant. Trattoria Spaghetto is an it-italian place. I mean, it provides... it offers really bad atmosphere.		
compare(Ikeno Hana, Menchanko-Tei)	**emotional stability**	**low**
I don't know! I might recommend Ikeno Hana and Menchanko-Tei. Menchanko-Tei features like, really nice food. It's err... in the Upper West Side and it's a japanese and noodles place, a ja-ja-japanese and noodles place. I mean, Ikeno Hana is in Midtown and a japanese and sushi place.		
compare(Kin Khao, Tossed)	**emotional stability**	**low**
I don't know. I might re-re-recommend Kin Khao and Tossed. Even if Kin Khao offers adequate food, I mean, it has bad waiters. Err... it's a thai place. Actually, Tossed is a cafe re-restaurant. It provides nice food. It features bad service. It has like, really rude waiters.		

Content plan	Trait	End
compare(Le Rivage, Pintaile's Pizza)	**emotional stability**	**low**
I don't know. I might ap-ap-approve Le Rivage and Pintaile's Pizza. Even if Pintaile's Pizza is cheap, it features like, bad atmosphere. Le Rivage is in Manhattan! Err... actually, it provides really bad atmosphere. Its price is 40 do-do-dollars.		
compare(Les Routiers, Radio Perfecto)	**emotional stability**	**low**
I don't know. I mean, actually, I might recommend Les Routiers and Radio Perfecto. Radio Perfecto's price is 25 dollars. It provides adequate food. It's err... in the East Village. Les Routiers is... it's expensive, but it features ac-ac-acceptable food th-though!		
compare(Mavalli Palace, Once Upon A Tart)	**emotional stability**	**low**
I don't know. I mean, actually, Mavalli Palace and Once Upon A Tart are the only places I would recommend! Err... Mavalli Palace's price is 27 dollars. This restaurant, which is an indian place, an in-indian place, is in Murray Hill, located in Murray Hill. Once Upon A Tart is in TriBeCa/SoHo.		
compare(Old San Juan, Veselka)	**emotional stability**	**low**
I might be wrong. Old San Juan is located in Ma-Manhattan. This restaurant, which features bad service, is a latin american place! I mean, actually, Veselka is located in Manhattan. Err... this eating place, which provides like, bad staff, is an east european place.		
compare(Scopa, Shabu, Tatsu)	**emotional stability**	**low**
I might be wrong. I might re-re-recommend Sc-Sc-Scopa and Shabu-Tatsu! Scopa has... it provides bad atmosphere, but it features like, nice service, though. It's an it-it-italian and new american place. I mean, Shabu-Tatsu is a japanese place. Even if the service is damn nice, it offers re-re-really bad atmosphere.		
compare(Soul Fixins', Walker's)	**emotional stability**	**low**
I don't know. Soul Fixins' and Walker's are the only places I would approve. Even if Walker's's price is 25 dollars, it provides like, bad service! Actually, I mean, it's a bar snacks place. Soul Fixins' is a southern place with really bad service.		
recommend(Cent anni)	**emotional stability**	**low**
I am not really sure. Cent'anni is the only restaurant I would recommend. It's an italian place. It offers bad at-at-atmosphere, but it features like, nice waiters, though. It provides good food. I mean, it's bloody expensive. Err... its price is 45 dollars.		
recommend(Chimichurri Grill)	**emotional stability**	**low**
I am not sure! I mean, Ch-Chimichurri Grill is the only place I would recommend. It's a latin american place. Err... its price is... it's damn ex-expensive, but it pr-pr-provides like, adequate food, though. It offers bad atmosphere, even if it features nice waiters.		
recommend(Edgar's Cafe)	**emotional stability**	**low**
I might be wrong. Ed-Ed-Edgar's Cafe is the only place I would approve. Although it's cheap, it features bad waiters. It provides like, really nice food! Err... it's a cafe restaurant. Actually, this restaurant, which offers poor ambience, is in Manhattan.		
recommend(Flor De Mayo)	**emotional stability**	**low**
I mean, I am not really sure. Fl-Flor De Mayo is the only place I would recommend. It's a chinese and latin american place! Ac-ac-actually, this restaurant, which features poor atmosphere although it provides like, nice food, is located in the Upper West Side with bad service.		
recommend(Jin Dal Lae)	**emotional stability**	**low**
I might be damn wrong. Jin Dal Lae is the only place I would recommend. It's in The Bronx! Ac-actually, it's a korean restaurant with bad service. I mean, it features... it provides like, poor ambience.		
recommend(John's Pizzeria)	**emotional stability**	**low**
I am not sure. I mean, John's Pizzeria is the only restaurant I would recommend. Even if it offers good food, it provides bad service. Err... it features like, really poor atmosphere, but it's bloody cheap though. It's an it-italian and pizza place and in Midtown.		
recommend(Pepolino)	**emotional stability**	**low**
I mean, I am not really sure. Pepolino is the only restaurant I would recommend. It provides poor atmosphere, but it offers friendly service though. Its price is 39 dollars, but it features like, tasty food, though. It's located in TriBeCa/SoHo. Err... actually, it's an italian place.		
recommend(Pietro s)	**emotional stability**	**low**
Ac-ac-actually, I might be wrong. Pietro's is the only place I would recommend. Err... it features nice waiters, it's in Mi-Midtown West and the atmosphere is damn bad. It's... although it provides like, really good food, its price is 56 dollars.		
recommend(Trattoria Rustica)	**emotional stability**	**low**
I am not sure. Trattoria Rustica is the only re-re-restaurant I would approve! Actually, this eating place, which is an italian place, is in Manhattan. Err... its price is 35 dollars, but it fe-fe-features friendly waiters though. It provides like, really poor at-atmosphere.		
recommend(Vinnie's Pizza)	**emotional stability**	**low**
I don't know. Ac-ac-actually, I would ap-ap-approve Vinnie's Pizza. It's a pizza place. Even if it's bloody cheap, I mean, it features like, really bad ambience. It provides rude staff.		
compare(Acacia, Marinella)	**emotional stability**	**high**
Let's see, Acacia and Marinella... I guess Acacia offers sort of decent food. Basically, Marinella, however, just has quite adequate food.		
compare(Caffe Cielo, Trattoria Spaghetto)	**emotional stability**	**high**
Let's see, Caffe Cielo and Trattoria Spaghetto... Caffe Cielo offers kind of acceptable food and Trattoria Spaghetto just has sort of satisfying food. Basically, I guess they're outstanding restaurants.		

Content plan	Trait	End
compare(Ikeno Hana, Menchanko-Tei)	**emotional stability**	**high**
Did you say Menchanko-Tei and Ikeno Hana? I think that Menchanko-Tei is a japanese and noodles restaurant. This eating house, which just has kind of passable food, is located in the Upper West Side. Basically, Ikeno Hana offers rather decent food.		
compare(Kin Khao, Tossed)	**emotional stability**	**high**
Let's see, Tossed and Kin Khao... Tossed, which offers adequate food, is a cafe place. However, basically, Kin Khao just has sort of decent food. It seems to me that they're kind of fantastic restaurants.		
compare(Le Rivage, Pintaile's Pizza)	**emotional stability**	**high**
Let's see, Le Rivage and Pintaile's Pizza... Le Rivage is in Manhattan. I think that Pintaile's Pizza, however, is in the Upper East Side, also its price is around 14 dollars. Basically, these eating places are quite phenomenal restaurants.		
compare(Les Routiers, Radio Perfecto)	**emotional stability**	**high**
You want to know more about Radio Perfecto and Les Routiers? Radio Perfecto offers sort of decent food and it's rather low-cost. However, Les Routiers just has adequate food. Basically, they're quite fantastic restaurants.		
compare(Mavalli Palace, Once Upon A Tart)	**emotional stability**	**high**
Let's see, Mavalli Palace and Once Upon A Tart... Mavalli Palace is just quite low-priced, Once Upon A Tart is in TriBeCa/SoHo and it's somewhat sort of inexpensive. These eating houses are kind of phenomenal restaurants.		
compare(Old San Juan, Veselka)	**emotional stability**	**high**
You want to know more about Veselka and Old San Juan? I think that Veselka, which is located in Manhattan, is an east european eating place. Basically, Old San Juan, however, is a latin american restaurant, also it's in Manhattan.		
compare(Scopa, Shabu, Tatsu)	**emotional stability**	**high**
You want to know more about Shabu-Tatsu and Scopa? Basically, I think that Shabu-Tatsu, which has rather acceptable waiters, is a japanese eating place. However, Scopa has quite satisfying staff. They're sort of phenomenal restaurants.		
compare(Soul Fixins', Walker's)	**emotional stability**	**high**
Let's see, Walker's and Soul Fixins'... Walker's is a bar snacks eating house and its price is 25 dollars. Basically, Soul Fixins', on the other hand, is somewhat quite inexpensive. I guess they're rather phenomenal restaurants.		
recommend(Cent anni)	**emotional stability**	**high**
Did you say Cent'anni? Basically, it seems to me that it's the best because the staff is somewhat quite adequate, also this eating house offers kind of tasty food.		
recommend(Chimichurri Grill)	**emotional stability**	**high**
Let's see what we can find on Chimichurri Grill. Basically, it's the best.		
recommend(Edgar's Cafe)	**emotional stability**	**high**
You want to know more about Edgar's Cafe? Basically, I think that Edgar's Cafe, which has rather decent food, is kind of the best restaurant.		
recommend(Flor De Mayo)	**emotional stability**	**high**
Did you say Flor De Mayo? Basically, it seems that Flor De Mayo, which just has rather satisfying food, is the best.		
recommend(Jin Dal Lae)	**emotional stability**	**high**
Let's see what we can find on Jin Dal Lae. Because Jin Dal Lae's price is around 27 dollars, also the food is somewhat quite satisfying just, this eating house is kind of the best restaurant.		
recommend(John's Pizzeria)	**emotional stability**	**high**
Did you say John's Pizzeria? Basically, it's kind of the best restaurant since the food is sort of tasty, also this eating place is somewhat quite inexpensive.		
recommend(Pepolino)	**emotional stability**	**high**
Did you say Pepolino? Basically, I think that Pepolino, which just has rather decent staff, is sort of the best restaurant of its kind.		
recommend(Pietro s)	**emotional stability**	**high**
You want to know more about Pietro's? Basically, Pietro's, which has kind of passable waiters, is sort of the best restaurant.		
recommend(Trattoria Rustica)	**emotional stability**	**high**
Let's see what we can find on Trattoria Rustica. Basically, it's the best because the service is just quite friendly.		
recommend(Vinnie's Pizza)	**emotional stability**	**high**
You want to know more about Vinnie's Pizza? It seems that it just has quite acceptable food. Basically, its price is around 13 dollars. This eating place is the best.		
compare(Acacia, Marinella)	**agreeableness**	**low**
Acacia and Marinella are damn alright. I mean, actually, Acacia doesn't provide nasty food. Obviously, it's located in Midtown. Basically, Marinella is in Manhattan.		
compare(Caffe Cielo, Trattoria Spaghetto)	**agreeableness**	**low**
Caffe Cielo and Trattoria Spaghetto are like, acceptable restaurants. Basically, I mean, Caffe Cielo provides decent food, also the atmosphere isn't good. Obviously, Trattoria Spaghetto offers bad ambiance. The ambience is bloody poor.		

Content plan	Trait	End
compare(Ikeno Hana, Menchanko-Tei)	**agreeableness**	**low**
Obviously, basically, Menchanko-Tei, which doesn't provide nasty food, is a japanese and noodles restaurant. I mean, actually, it's in the Upper West Side, this restaurant is located in the Upper West Side. Ikeno Hana is located in Midtown.		
compare(Kin Khao, Tossed)	**agreeableness**	**low**
Even if Tossed doesn't have nasty food, actually, the service is damn unmannered. I mean, basically, Kin Khao offers like, rude staff.		
compare(Le Rivage, Pintaile's Pizza)	**agreeableness**	**low**
Actually, there could be worse places. Pintaile's Pizza's price is 14 dollars. Basically, Le Rivage provides like, mediocre ambiance. I mean, it isn't cheap, this restaurant's price is 40 dollars.		
compare(Les Routiers, Radio Perfecto)	**agreeableness**	**low**
Actually, there could be worse restaurants. Radio Perfecto doesn't offer nasty food. It's in the East Village. Basically, Les Routiers is in Manhattan. I mean, its price is 41 dollars.		
compare(Mavalli Palace, Once Upon A Tart)	**agreeableness**	**low**
Basically, there could be worse places. Actually, Mavalli Palace's price is 27 dollars. It's an indian restaurant. I mean, Once Upon A Tart, which is in TriBeCa/SoHo, is a cafe and sandwich place, a cafe and sandwich place.		
compare(Old San Juan, Veselka)	**agreeableness**	**low**
Actually, I mean, there could be worse places. Old San Juan provides like, bad staff. Veselka is in Manhattan. Basically, the waiters aren't good.		
compare(Scopa, Shabu, Tatsu)	**agreeableness**	**low**
Everybody knows that there could be worse restaurants. Scopa doesn't provide good atmosphere. Basically, Shabu-Tatsu has like, decent waiters. I mean, the ambience is damn poor. This restaurant offers mediocre ambiance.		
compare(Soul Fixins', Walker's)	**agreeableness**	**low**
Actually, I mean, there could be worse places. Even if Walker's is low-cost, the service isn't good. Soul Fixins' is a southern restaurant. Basically, the staff is rude.		
recommend(Cent anni)	**agreeableness**	**low**
Basically, Cent'anni is the only place that is any good. This restaurant is located in Manhattan. Obviously, it's an italian restaurant. This restaurant's price is 45 dollars, although the waiters aren't unfriendly. This restaurant has like, poor ambience. Actually, the ambiance is bloody mediocre.		
recommend(Chimichurri Grill)	**agreeableness**	**low**
I mean, Chimichurri Grill isn't as bad as the others. Basically, the staff isn't nasty. Actually, its price is 41 dollars. It's damn costly.		
recommend(Edgar's Cafe)	**agreeableness**	**low**
Basically, I mean, I thought everybody knew that Edgar's Cafe is the only restaurant that is any good. Actually, it isn't expensive.		
recommend(Flor De Mayo)	**agreeableness**	**low**
Basically, Flor De Mayo isn't as bad as the others. Obviously, it isn't expensive. I mean, actually, its price is 18 dollars.		
recommend(Jin Dal Lae)	**agreeableness**	**low**
There could be worse places. Jin Dal Lae is located in The Bronx. I mean, basically, the staff is rude. Actually, this restaurant doesn't provide good ambiance.		
recommend(John's Pizzeria)	**agreeableness**	**low**
Basically, everybody knows that John's Pizzeria is the only restaurant that is any good. This eating place offers like, mediocre ambiance, even if it's bloody inexpensive. Actually, the waiters aren't good.		
recommend(Pepolino)	**agreeableness**	**low**
I mean, I thought everybody knew that Pepolino is the only place that is any good. Actually, this restaurant is in TriBeCa/SoHo. Basically, this restaurant, which doesn't provide good ambiance, is damn pricy.		
recommend(Pietro s)	**agreeableness**	**low**
Everybody knows that Pietro's isn't as bad as the others. Actually, the food isn't nasty. I mean, basically, this restaurant is bloody pricy.		
recommend(Trattoria Rustica)	**agreeableness**	**low**
I thought everybody knew that Trattoria Rustica isn't as bad as the others. Although the food isn't nasty, it's pricy. I mean, actually, it's in Manhattan. Although the waiters are bloody adequate, basically, the ambience is poor.		
recommend(Vinnie's Pizza)	**agreeableness**	**low**
Vinnie's Pizza is the only place that is any good. Obviously, basically, it's damn cheap. This restaurant has like, rude waiters. Actually, the staff isn't good. This restaurant, which is in Manhattan, is a pizza place.		
compare(Acacia, Marinella)	**agreeableness**	**high**
Did you say Acacia and Marinella? I imagine you would appreciate them, you see? It seems to me that Marinella provides kind of satisfactory food, also it's an italian place mate, but Acacia offers sort of acceptable food, you know.		

Content plan	Trait	End
compare(Caffe Cielo, Trattoria Spaghetto)	**agreeableness**	**high**
Let's see what we can find on Caffe Cielo and Trattoria Spaghetto. Oh yeah, I would approve them. Even if Caffe Cielo has rather bad atmosphere, it offers sort of nice food, alright? I guess Trattoria Spaghetto is an italian place. Even if it provides bad atmosphere, this restaurant features nice food.		
compare(Ikeno Hana, Menchanko-Tei)	**agreeableness**	**high**
Let's see what we can find on Ikeno Hana and Menchanko-Tei. Well, yeah, I would advise them. I suppose Ikeno Hana provides quite nice food, you know. However, Menchanko-Tei, which offers nice food, is a japanese and noodles place, okay?		
compare(Kin Khao, Tossed)	**agreeableness**	**high**
Did you say Kin Khao and Tossed? Oh yeah, I would suggest them, wouldn't you? Kin Khao offers quite satisfactory food. I guess Tossed, however, has sort of acceptable food.		
compare(Le Rivage, Pintaile's Pizza)	**agreeableness**	**high**
You want to know more about Le Rivage and Pintaile's Pizza? I see, right, I would suggest them, you would probably like them, you see? It seems to me that Pintaile's Pizza is inexpensive. Le Rivage is located in Manhattan and it offers quite bad atmosphere, also it's somewhat expensive.		
compare(Les Routiers, Radio Perfecto)	**agreeableness**	**high**
Let's see, Les Routiers and Radio Perfecto... Oh i see, I would consider them, I suspect you would enjoy them, alright? Les Routiers, which offers quite acceptable food, is in Manhattan. It seems to me that Radio Perfecto is sort of affordable. This restaurant, which has adequate food, is in the East Village.		
compare(Mavalli Palace, Once Upon A Tart)	**agreeableness**	**high**
You want to know more about Mavalli Palace and Once Upon A Tart? Well, yeah, I would advise them. I suppose Once Upon A Tart's price is around 15 dollars, it's located in TriBeCa/SoHo, also it's a cafe and sandwich place, it's a cafe and sandwich place, Mavalli Palace's price is 27 dollars, it's kind of affordable, and it's located in Murray Hill and an indian place, okay?		
compare(Old San Juan, Veselka)	**agreeableness**	**high**
Let's see, Old San Juan and Veselka... Oh right, I would suggest them. I suppose Veselka is an east european place and located in Manhattan and Old San Juan is a latin american place and in Manhattan pal, okay? you know.		
compare(Scopa, Shabu, Tatsu)	**agreeableness**	**high**
Let's see, Scopa and Shabu-Tatsu... It seems that Scopa is an italian and new american place with rather nice service while Shabu-Tatsu is a japanese place with quite nice waiters, you know pal, okay?		
compare(Soul Fixins', Walker's)	**agreeableness**	**high**
You want to know more about Soul Fixins' and Walker's? Yeah, I would suggest them, I guess they're kind of alright, okay? Soul Fixins' is somewhat quite cheap, also it's a southern place. It seems that Walker's's price is 25 dollars, also it's a bar snacks place.		
recommend(Cent anni)	**agreeableness**	**high**
Did you say Cent'anni? I imagine you would appreciate it, you see? It seems that this eating place, which provides sort of good food and rather acceptable service, you know, is in Manhattan mate.		
recommend(Chimichurri Grill)	**agreeableness**	**high**
You want to know more about Chimichurri Grill? I guess you would like it buddy because this restaurant, which is in Midtown West, is a latin american place with rather nice food and quite nice waiters, you know, okay?		
recommend(Edgar's Cafe)	**agreeableness**	**high**
Let's see, Edgar's Cafe... Right, well, it's one of my favourite restaurants. I guess it's somewhat inexpensive. This eating house, which offers rather acceptable food, is a cafe eating place and located in Manhattan, okay?		
recommend(Flor De Mayo)	**agreeableness**	**high**
Let's see, Flor De Mayo... I see, well, it's one of my favourite places, isn't it? I think that its price is 18 dollars, you know. This restaurant, which is in the Upper West Side, is a chinese and latin american place with sort of nice food.		
recommend(Jin Dal Lae)	**agreeableness**	**high**
You want to know more about Jin Dal Lae? Right, yeah, I think that it's in The Bronx with kind of acceptable food, so I guess you would like it, I would suggest this restaurant, and its price is around 27 dollars, also this restaurant is a korean place, you see?		
recommend(John's Pizzeria)	**agreeableness**	**high**
You want to know more about John's Pizzeria? Oh well, it's price is around 20 dollars, so it's one of my favourite eating houses, and it offers kind of dainty food, you know, you see?		
recommend(Pepolino)	**agreeableness**	**high**
Let's see, Pepolino... I imagine you would appreciate it buddy since it's in TriBeCa/SoHo and an italian place with rather nice waiters, you know, okay?		
recommend(Pietro s)	**agreeableness**	**high**
You want to know more about Pietro's? I believe you would love it, you know, alright? It seems to me that it features rather good food. This eating place, which provides quite acceptable service, is an italian and steak house eating house and in Midtown West mate.		
recommend(Trattoria Rustica)	**agreeableness**	**high**
You want to know more about Trattoria Rustica? I imagine you would appreciate it, you know pal, alright? It offers nice food, even if this restaurant is somewhat expensive. This restaurant, which features quite friendly waiters although this restaurant provides kind of bad atmosphere, is an italian place.		

Content plan	Trait	End
recommend(Vinnie's Pizza)	**agreeableness**	**high**
Let's see, Vinnie's Pizza... I imagine you would appreciate it, it's one of my favourite places, you know pal, you see?		
compare(Acacia, Marinella)	**conscientiousness**	**low**
Err... I am not sure. Acacia and Marinella are damn alright! Mmhm... Acacia provides kind of nice food, it's a bistro and french place and Marinella is an italian place and located in Manhattan.		
compare(Caffe Cielo, Trattoria Spaghetto)	**conscientiousness**	**low**
I might be wrong! I mean, there could be worse places. Err... Caffe Cielo is an italian place. Even if it offers like, nice food, the atmosphere isn't good. Trattoria Spaghetto is an italian place and the atmosphere is damn bad, this restaurant provides bad atmosphere.		
compare(Ikeno Hana, Menchanko-Tei)	**conscientiousness**	**low**
I might be damn wrong. I mean, I might suggest Ikeno Hana and Menchanko-Tei. Mmhm... Menchanko-Tei offers like, nice food, Ikeno Hana is a japanese and sushi place, it's a japanese and sushi place, and this restaurant is located in Midtown!		
compare(Kin Khao, Tossed)	**conscientiousness**	**low**
I might be bloody wrong. Kin Khao is a thai place, also the waiters aren't good. Mmhm... even if Tossed provides kind of nice food mate, it offers like, bad staff, bad waiters.		
compare(Le Rivage, Pintaile's Pizza)	**conscientiousness**	**low**
I am not sure. There could be worse places! Mmhm... even if Pintaile's Pizza isn't kind of expensive, it offers like, bad ambience, bad ambiance, but Le Rivage is bloody costly.		
compare(Les Routiers, Radio Perfecto)	**conscientiousness**	**low**
Mmhm... err... I might be kind of wrong. Acceptable places. I mean, Radio Perfecto is located in the East Village with nice food, Les Routiers is located in Manhattan and its price is 41 dollars, this restaurant isn't cheap!		
compare(Mavalli Palace, Once Upon A Tart)	**conscientiousness**	**low**
I might be darn wrong! Err... I mean, I might advise Mavalli Palace and Once Upon A Tart. Mmhm... Mavalli Palace's price is 27 dollars and Once Upon A Tart is a cafe and sandwich place, it's a cafe and sandwich place, also it's in TriBeCa/SoHo.		
compare(Old San Juan, Veselka)	**conscientiousness**	**low**
Mmhm... err... I might be bloody wrong. Old San Juan is a latin american place and in Manhattan, also the waiters aren't kind of good, and Veselka provides like, bad service.		
compare(Scopa, Shabu, Tatsu)	**conscientiousness**	**low**
I might be darn wrong. Scopa provides kind of bad ambiance! Err... even if Shabu-Tatsu doesn't have unfriendly waiters, it offers like, poor ambience, bad atmosphere.		
compare(Soul Fixins', Walker's)	**conscientiousness**	**low**
I don't know mate! Even if Walker's's price is 25 dollars, the service isn't good. This restaurant is a bar snacks place. Err... Soul Fixins' is a southern place. The service is bloody bad, it features like, bad waiters.		
recommend(Cent anni)	**conscientiousness**	**low**
I am not kind of sure! I mean, Cent'anni's price is 45 dollars, so this restaurant is the only place that is any good, it's damn expensive, this restaurant has nice waiters though mate and the atmosphere isn't good.		
recommend(Chimichurri Grill)	**conscientiousness**	**low**
I am not kind of sure pal. Err... Chimichurri Grill is the only place I would advise. It doesn't provide unfriendly service! This restaurant is damn expensive, its price is 41 dollars.		
recommend(Edgar's Cafe)	**conscientiousness**	**low**
I might be bloody wrong. Err... Edgar's Cafe is the only place that is any good. I mean, this restaurant offers like, bad service, bad waiters, even if the food isn't kind of nasty.		
recommend(Flor De Mayo)	**conscientiousness**	**low**
I mean, I am not sure. Flor De Mayo isn't as bad as the others buddy. Mmhm... even if it isn't expensive, this restaurant features like, bad waiters, bad service. Err... it's a chinese and latin american place and located in the Upper West Side.		
recommend(Jin Dal Lae)	**conscientiousness**	**low**
Err... mmhm... I don't know buddy. There could be worse places. I mean, Jin Dal Lae provides kind of bad atmosphere, poor atmosphere, even if its price is 27 dollars.		
recommend(John's Pizzeria)	**conscientiousness**	**low**
Mmhm... I might be wrong. John's Pizzeria is the only place that is any good. This restaurant provides poor atmosphere, bad atmosphere, even if the food is bloody good. Err... this restaurant doesn't feature kind of good service!		
recommend(Pepolino)	**conscientiousness**	**low**
Err... I don't know. I mean, Pepolino is the only place that is any good. Mmhm... this restaurant is an italian place and in TriBeCa/SoHo with kind of bad atmosphere, also its price is 39 dollars, this restaurant isn't cheap!		
recommend(Pietro s)	**conscientiousness**	**low**
Err... I am not kind of sure mate. I mean, Pietro's is an italian and steak house place and in Midtown West, so I would approve this restaurant, and the atmosphere is darn bad, also this restaurant's price is 56 dollars, it isn't cheap.		

Content plan	Trait	End
recommend(Trattoria Rustica)	**conscientiousness**	**low**
Err... mmhm... I am not sure. I mean, Trattoria Rustica is the only place that is any good! This restaurant features kind of nice waiters. It isn't cheap. Its price is 35 dollars.		
recommend(Vinnie's Pizza)	**conscientiousness**	**low**
I might be wrong! Vinnie's Pizza is err... in Manhattan and a pizza place, so this restaurant isn't as bad as the others, mate and it features like, kind of bad waiters, with bad atmosphere.		
compare(Acacia, Marinella)	**conscientiousness**	**high**
You want to know more about Acacia and Marinella? I imagine you would appreciate them. I think that Marinella, which offers quite satisfactory food, is in Manhattan and an italian eating house. However, Acacia provides sort of satisfying food and it's located in Midtown.		
compare(Caffe Cielo, Trattoria Spaghetto)	**conscientiousness**	**high**
Let's see, Caffe Cielo and Trattoria Spaghetto... They're rather outstanding. I guess Trattoria Spaghetto offers sort of acceptable food, also it's an italian eating house. Caffe Cielo, which provides quite satisfactory food, is an italian eating place.		
compare(Ikeno Hana, Menchanko-Tei)	**conscientiousness**	**high**
Did you say Ikeno Hana and Menchanko-Tei? I see, well, I suppose they're quite phenomenal eating houses. Ikeno Hana provides sort of satisfying food. Menchanko-Tei, which offers satisfactory food, on the other hand, is a japanese and noodles eating place.		
compare(Kin Khao, Tossed)	**conscientiousness**	**high**
You want to know more about Kin Khao and Tossed? Well, i see, I suppose Kin Khao, which provides sort of satisfying food, is a thai eating place. However, Tossed, which offers rather satisfactory food, is a cafe eating house.		
compare(Le Rivage, Pintaile's Pizza)	**conscientiousness**	**high**
Let's see, Le Rivage and Pintaile's Pizza... I see, well, they're rather outstanding eating houses. Even if Pintaile's Pizza provides quite unspeakable atmosphere, it's sort of inexpensive. On the other hand, Le Rivage is in Manhattan, also its price is 40 dollars.		
compare(Les Routiers, Radio Perfecto)	**conscientiousness**	**high**
Let's see what we can find on Les Routiers and Radio Perfecto. Well, i see, they're outstanding eating places. It seems to me that Les Routiers offers satisfactory food. Radio Perfecto is somewhat low-priced. This eating house, which provides quite satisfying food, is in the East Village.		
compare(Mavalli Palace, Once Upon A Tart)	**conscientiousness**	**high**
Did you say Mavalli Palace and Once Upon A Tart? I see, well, they're rather phenomenal eating places. I think that Once Upon A Tart is somewhat inexpensive. Mavalli Palace's price is 27 dollars. It's an indian eating house.		
compare(Old San Juan, Veselka)	**conscientiousness**	**high**
Let's see, Old San Juan and Veselka... I believe you would love them. I guess Old San Juan is a latin american eating place and in Manhattan. However, Veselka is an east european eating house and in Manhattan.		
compare(Scopa, Shabu, Tatsu)	**conscientiousness**	**high**
You want to know more about Scopa and Shabu-Tatsu? I see, well, they're quite outstanding eating places. I suppose Scopa is an italian and new american eating house with satisfying service. Shabu-Tatsu, which provides sort of satisfactory service, however, is a japanese restaurant.		
compare(Soul Fixins', Walker's)	**conscientiousness**	**high**
Let's see, Soul Fixins' and Walker's... I see, they're sort of outstanding eating houses. Even if Soul Fixins' provides unmannerly service, it's quite inexpensive. This eating place is a southern restaurant. It seems that Walker's is somewhat low-priced, even if the service is ill-mannered.		
recommend(Cent anni)	**conscientiousness**	**high**
Did you say Cent'anni? Well, i see, it's quite the best eating house since it offers rather dainty food with sort of satisfactory service.		
recommend(Chimichurri Grill)	**conscientiousness**	**high**
Let's see what we can find on Chimichurri Grill. I guess you would like it since this eating house, which offers sort of satisfying food and quite satisfactory waiters, is a latin american eating place.		
recommend(Edgar's Cafe)	**conscientiousness**	**high**
Did you say Edgar's Cafe? I imagine you would appreciate it. I guess it's sort of inexpensive and the food is somewhat quite satisfactory, also it's a cafe eating place.		
recommend(Flor De Mayo)	**conscientiousness**	**high**
Let's see, Flor De Mayo... I imagine you would appreciate it since its price is 18 dollars.		
recommend(Jin Dal Lae)	**conscientiousness**	**high**
Let's see, Jin Dal Lae... I guess you would like it. I think that this eating house, which offers rather satisfactory food, is a korean eating place. Its price is 27 dollars.		
recommend(John's Pizzeria)	**conscientiousness**	**high**
Let's see what we can find on John's Pizzeria. I see, well, it's one of my favourite eating houses. It offers dainty food, although the service is ill-mannered. It seems that this eating place, which is located in Midtown, is somewhat inexpensive, even if the atmosphere is rather second-rate.		
recommend(Pepolino)	**conscientiousness**	**high**
Let's see, Pepolino... I guess you would like it since this eating place, which offers rather satisfying staff, is located in TriBeCa/SoHo with sort of dainty food.		
recommend(Pietro s)	**conscientiousness**	**high**
Let's see, Pietro's... I guess you would like it since the food is sort of dainty with quite satisfying staff.		

<table>
<tr><th>Content plan</th><th>Trait</th><th>End</th></tr>
<tr><td>recommend(Trattoria Rustica)</td><td>conscientiousness</td><td>high</td></tr>
<tr><td colspan="3">Let's see what we can find on Trattoria Rustica. Well, i see, Trattoria Rustica, which is in Manhattan and an italian eating place with rather satisfactory food and sort of satisfying staff, is quite the best eating house.</td></tr>
<tr><td>recommend(Vinnie's Pizza)</td><td>conscientiousness</td><td>high</td></tr>
<tr><td colspan="3">You want to know more about Vinnie's Pizza? I see, it's one of my favourite eating houses. It's somewhat sort of inexpensive. I suppose this restaurant provides quite satisfactory food. This eating place, which is a pizza restaurant, is in Manhattan.</td></tr>
<tr><td>compare(Acacia, Marinella)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">I mean, I might be darn wrong. Mmhm... err... I might suggest Acacia and Marinella. Acacia offers like, nice food, nice food, and Marinella is located in Manhattan.</td></tr>
<tr><td>compare(Caffe Cielo, Trattoria Spaghetto)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Mmhm... I mean, I am not sure. I might suggest Caffe Cielo and Trattoria Spaghetto. Err... Caffe Cielo provides nice food, it has bad atmosphere and Trattoria Spaghetto is an italian place, also it features... it offers like, poor atmosphere.</td></tr>
<tr><td>compare(Ikeno Hana, Menchanko-Tei)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">I mean, I might be darn wrong. Mmhm... I might consider Ikeno Hana and Menchanko-Tei. Err... Menchanko-Tei has like, nice food, also it's... it's in the Upper West Side, and Ikeno Hana is located in Midtown.</td></tr>
<tr><td>compare(Kin Khao, Tossed)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Err... I don't know. Mmhm... I mean, Kin Khao is a thai place, it features bad waiters and Tossed has like, nice food, and bad waiters, rude staff.</td></tr>
<tr><td>compare(Le Rivage, Pintaile's Pizza)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Mmhm... err... I don't know. I mean, Pintaile's Pizza's price is 14 dollars, also it has like, bad atmosphere, bad atmosphere, and Le Rivage is darn expensive.</td></tr>
<tr><td>compare(Les Routiers, Radio Perfecto)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Mmhm... I don't know. Err... I mean, Radio Perfecto offers like, nice food, it's in the East Village and Les Routiers is... its price is 41 dollars.</td></tr>
<tr><td>compare(Mavalli Palace, Once Upon A Tart)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Err... I am not sure. Mmhm... I mean, Mavalli Palace is low-cost and Once Upon A Tart is a cafe and sandwich place, it's a cafe and sandwich place, also it's located in TriBeCa/SoHo.</td></tr>
<tr><td>compare(Old San Juan, Veselka)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">I don't know. Mmhm... I mean, I might consider Old San Juan and Veselka. Err... Old San Juan is a latin american place, also it offers bad service, and Veselka has... it provides like, bad staff.</td></tr>
<tr><td>compare(Scopa, Shabu, Tatsu)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Err... I don't know. Mmhm... I might consider Scopa and Shabu-Tatsu. I mean, Scopa offers like, bad ambience, Shabu-Tatsu has nice service and the atmosphere is poor.</td></tr>
<tr><td>compare(Soul Fixins', Walker's)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">I don't know. Acceptable places. Even if Walker's is low-cost, I mean, it has bad waiters. Mmhm... err... Soul Fixins' has like, bad waiters, it offers bad service.</td></tr>
<tr><td>recommend(Cent anni)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Mmhm... err... I don't know. I mean, I would consider Cent'anni because it has like, good food.</td></tr>
<tr><td>recommend(Chimichurri Grill)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Err... I am not sure. Mmhm... I mean, Chimichurri Grill offers like, nice food, so I would advise it, also the atmosphere is bad and its price is 41 dollars.</td></tr>
<tr><td>recommend(Edgar's Cafe)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Err... I mean, I might be darn wrong. Mmhm... because Edgar's Cafe is... it's in Manhattan, you would probably like it.</td></tr>
<tr><td>recommend(Flor De Mayo)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Err... I mean, I am not sure. Mmhm... Flor De Mayo offers... so I would advise it, it provides like, nice food.</td></tr>
<tr><td>recommend(Jin Dal Lae)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Err... I mean, I might be darn wrong. Jin Dal Lae is mmhm... in The Bronx, so I would consider it, and it has like, bad atmosphere.</td></tr>
<tr><td>recommend(John's Pizzeria)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">I mean, I am not sure. Mmhm... John's Pizzeria is err... in Midtown, so I would suggest it, it's located in Midtown.</td></tr>
<tr><td>recommend(Pepolino)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Mmhm... I am not sure. Err... I mean, Pepolino has... so I would suggest it, it offers like, nice service.</td></tr>
<tr><td>recommend(Pietro s)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Mmhm... err... I might be wrong. I mean, I would consider Pietro's because it has like, nice staff, also its price is... it's darn expensive.</td></tr>
<tr><td>recommend(Trattoria Rustica)</td><td>openness to experience</td><td>low</td></tr>
<tr><td colspan="3">Err... I don't know. Mmhm... I mean, Trattoria Rustica is an italian place, so I would approve it, and it's expensive.</td></tr>
</table>

Content plan	Trait	End
recommend(Vinnie's Pizza)	**openness to experience**	**low**
Err... I mean, I am not sure. I would suggest... it isn't as bad as the others.		
compare(Acacia, Marinella)	**openness to experience**	**high**
You want to know more about Acacia and Marinella? It seems to me that they're kind of outstanding, you know. Acacia, which features rather satisfactory food, is a bistro and french eating house and in Midtown while Marinella is an italian eating place and in Manhattan, you see?		
compare(Caffe Cielo, Trattoria Spaghetto)	**openness to experience**	**high**
You want to know more about Caffe Cielo and Trattoria Spaghetto? I think that they're rather phenomenal eating places. Caffe Cielo, which provides kind of satisfactory food even if the atmosphere is somewhat second-rate, you know, is an italian restaurant. Trattoria Spaghetto, which features acceptable food although this eating house offers mediocre atmosphere, is an italian place, you see?		
compare(Ikeno Hana, Menchanko-Tei)	**openness to experience**	**high**
Let's see, Ikeno Hana and Menchanko-Tei... They're sort of outstanding. I think that Menchanko-Tei features kind of satisfactory food and it's a japanese and noodles restaurant, also this eating place is in the Upper West Side, you know, you see? On the other hand, Ikeno Hana is a japanese and sushi eating house.		
compare(Kin Khao, Tossed)	**openness to experience**	**high**
Let's see, Kin Khao and Tossed... They're phenomenal restaurants, aren't they? It seems to me that Kin Khao, which features rather ill-mannered service, is a thai eating house, you know. Tossed is a cafe eating place. Although the staff is somewhat sort of unmannerly, this restaurant provides satisfactory food.		
compare(Le Rivage, Pintaile's Pizza)	**openness to experience**	**high**
Did you say Le Rivage and Pintaile's Pizza? They're sort of outstanding restaurants, you know. Pintaile's Pizza's price is around 14 dollars, isn't it? It's located in the Upper East Side. I suppose Le Rivage is located in Manhattan. The ambiance is somewhat mediocre. This eating place's price is 40 dollars.		
compare(Les Routiers, Radio Perfecto)	**openness to experience**	**high**
Let's see, Les Routiers and Radio Perfecto... I see, they're phenomenal eating houses, aren't they? I guess Radio Perfecto is somewhat kind of low-priced and the food is sort of satisfactory, also it's in the East Village. Although Les Routiers's price is 41 dollars, the food, however, is satisfying.		
compare(Mavalli Palace, Once Upon A Tart)	**openness to experience**	**high**
Did you say Mavalli Palace and Once Upon A Tart? I see, it seems to me that Mavalli Palace's price is around 27 dollars, also it's located in Murray Hill and an indian eating place, you know and Once Upon A Tart is located in TriBeCa/SoHo, also it's a cafe and sandwich eating house, you see?		
compare(Old San Juan, Veselka)	**openness to experience**	**high**
Let's see, Old San Juan and Veselka... I see, they're quite phenomenal eating houses, you know, alright? Old San Juan, which provides kind of unmannered staff, is in Manhattan and a latin american restaurant. Veselka, which features sort of unmannerly waiters, is in Manhattan.		
compare(Scopa, Shabu, Tatsu)	**openness to experience**	**high**
You want to know more about Scopa and Shabu-Tatsu? I see, you would probably love them. I think that Scopa, which features kind of second-rate atmosphere, is an italian and new american eating place. Shabu-Tatsu, which offers sort of satisfactory service even if this restaurant provides quite mediocre atmosphere, is a japanese eating house, okay?		
compare(Soul Fixins', Walker's)	**openness to experience**	**high**
You want to know more about Walker's and Soul Fixins'? I see, Walker's, which is a bar snacks eating place, is rather low-priced, even if this eating house features sort of unmannered waiters okay? It seems that Soul Fixins' is a southern restaurant and the waiters are somewhat ill-mannered.		
recommend(Cent anni)	**openness to experience**	**high**
Let's see, Cent'anni... I see, it's quite the best restaurant, isn't it? This eating place, which is in Manhattan, is an italian eating house. This restaurant, which provides rather second-rate atmosphere, is somewhat sort of high-priced.		
recommend(Chimichurri Grill)	**openness to experience**	**high**
You want to know more about Chimichurri Grill? I believe you would love it, you know. I guess it's in Midtown West. Although this eating house's price is around 41 dollars, the food is rather satisfactory. This eating place, which provides kind of second-rate atmosphere, is a latin american restaurant, alright?		
recommend(Edgar's Cafe)	**openness to experience**	**high**
You want to know more about Edgar's Cafe? I see, it's the best. I think that its price is around 19 dollars, although this eating place features kind of ill-mannered service alright? This eating house, which provides quite second-rate atmosphere, is a cafe restaurant.		
recommend(Flor De Mayo)	**openness to experience**	**high**
Let's see what we can find on Flor De Mayo. I believe you would love it. It's a chinese and latin american eating house, you know. I guess this restaurant, which provides quite ill-mannered staff, is somewhat inexpensive, although this eating place features rather second-rate atmosphere okay?		

Content plan	Trait	End
recommend(Jin Dal Lae)	**openness to experience**	**high**
You want to know more about Jin Dal Lae? I believe you would love it. It seems to me that its price is around 27 dollars, although the waiters are rather unmannered. This restaurant is a korean eating house, isn't it? Although the ambiance is somewhat second-rate, this eating place provides sort of satisfying food.		
recommend(John's Pizzeria)	**openness to experience**	**high**
Let's see what we can find on John's Pizzeria. I believe you would love it. I suppose it's an italian and pizza eating place, you know, okay? This eating house, which features quite ill-mannered waiters and rather second-rate atmosphere, is in Midtown.		
recommend(Pepolino)	**openness to experience**	**high**
Let's see, Pepolino... It seems that it's the only eating house that is any good, you know. Although the ambiance is mediocre, the waiters are somewhat sort of satisfying. This restaurant, which provides dainty food although this eating place is quite high-priced, is an italian restaurant, okay?		
recommend(Pietro s)	**openness to experience**	**high**
You want to know more about Pietro's? I imagine you would appreciate it. The waiters are kind of satisfactory, although this restaurant provides sort of second-rate atmosphere alright? I suppose this eating place, which offers dainty food although this eating house's price is around 56 dollars, is in Midtown West, you know.		
recommend(Trattoria Rustica)	**openness to experience**	**high**
You want to know more about Trattoria Rustica? I see, it's the best place of its kind. Even if it's somewhat rather high-priced, this eating place provides sort of satisfactory food, you see? It's an italian restaurant. This eating house, which features satisfying waiters although the atmosphere is quite second-rate, is located in Manhattan.		
recommend(Vinnie's Pizza)	**openness to experience**	**high**
You want to know more about Vinnie's Pizza? I see, it's sort of the best restaurant. It's located in Manhattan, isn't it? Although this eating place features unmannerly waiters, the food is somewhat kind of satisfying. Its price is 13 dollars. I suppose this eating house provides terrible ambience.		

Appendix B

Utterances Generated using Random Parameters

This appendix contains a subset of the utterances generated by varying the input parameter values with a uniform distribution. These utterances were used for the feature analysis in Chapter 7, as well as for learning the data-driven models presented in Chapters 8 and 9.

compare(Kin Khao, Tossed) Kin Khao and Tossed are bloody outstanding. Kin Khao just has rude staff. Tossed features sort of unmannered waiters, even if the food is somewhat quite adequate.
compare(Kin Khao, Tossed) Err... I am not really sure. Tossed offers kind of decent food. Mmhm... however, Kin Khao, which has quite ad-ad-adequate food, is a thai place. You would probably enjoy these restaurants.
compare(Kin Khao, Tossed) Mmhm... well, Tossed is a cafe restaurant with really satisfactory food, you know. Err... on the other hand, Kin Khao, which doesn't provide nasty food, is a thai eating place, you see?
compare(Kin Khao, Tossed) Yeah, right, I might advise Kin Khao and Tossed. Err... Kin Khao, which features bad service even if this restaurant provides adequate food, is a thai place, you know. I guess Tossed has bad waiters, but it offers nice food though.
compare(Le Rivage, Pintaile's Pizza) Ok, right, Pintaile's Pizza is rather cheap. The atmosphere is bad, isn't it? I guess Le Rivage is in Manhattan, also its price is... it's expensive, you know.
compare(Le Rivage, Pintaile's Pizza) Yeah, I thought everybody knew that Le Rivage's price is around 40 dollars, it's located in Manhattan with kind of poor atmosphere, Pintaile's Pizza is located in the Upper East Side and its price is 14 dollars. They're... I would advise these restaurants, alright?
compare(Le Rivage, Pintaile's Pizza) Let's see, Pintaile's Pizza and Le Rivage... There could be worse restaurants. I mean, Pintaile's Pizza's price is 14 dollars. Err... Le Rivage, on the other hand, offers like, kind of poor ambience, and it's quite pricy.
compare(Le Rivage, Pintaile's Pizza) Oh Le Rivage is somewhat high-priced, also it has like, second-rate atmosphere, and this restaurant is in Manhattan. Pi-Pintaile's Pizza, which isn't expensive, is located in the Upper East Side. Basically, there... be worse places, these eating places are quite satisfactory eating houses.
compare(Le Rivage, Pintaile's Pizza) Ok, right, I thought everybody knew that Le Rivage and Pintaile's Pizza are... you would probably love them, okay? Pintaile's Pizza, which is in the Upper East Side, is rather cheap, inexpensive, while Le Rivage is in Manhattan and its price is 40 dollars.
compare(Le Rivage, Pintaile's Pizza) Mmhm... basically, Le Rivage's price is 40 dollars and it's in Manhattan with kind of poor atmosphere. Although Pintaile's Pizza's price is 14 dollars, actually, it provides like, bad atmosphere. This restaurant is in the Upper East Side.
compare(Le Rivage, Pintaile's Pizza) Obviously, ok, I might recommend Le Rivage and Pintaile's Pizza. Actually, I suppose Pintaile's Pizza's price is 14 dollars. Err... on the other hand, Le Rivage's price is 40 dollars.
compare(Le Rivage, Pintaile's Pizza) Actually, I am not sure mate. I might consider Le Rivage and Pintaile's Pizza. Pintaile's Pizza is damn cheap, Le Rivage is in Manhattan and the ambience is quite poor, also it's rather expensive, you see?
compare(Lemongrass Grill, Monsoon) Basically, Lemongrass Grill just has kind of second-rate atmosphere, the ambiance is mediocre, it's a thai restaurant and Monsoon doesn't feature nasty ambience!
compare(Lemongrass Grill, Monsoon) Well, actually, Monsoon has pleasant ambience, quite acceptable ambiance, but Lemongrass Grill, which doesn't feature good atmosphere, is a thai restaurant!
compare(Lemongrass Grill, Monsoon) Monsoon has sort of acceptable ambience. However, actually, Lemongrass Grill features mediocre ambiance.
compare(Lemongrass Grill, Monsoon) Actually, Lemongrass Grill is a thai restaurant, you know. Actually, Monsoon, which has acceptable ambience, however, is a vietnamese place, you know.
compare(Lemongrass Grill, Monsoon) Actually, Monsoon features nice atmosphere while Lemongrass Grill is a thai place and it doesn't have good ambience!
compare(Lemongrass Grill, Monsoon) Lemongrass Grill has poor ambience. Monsoon is a really vietnamese place, it's a vietnamese restaurant. The atmosphere is really nice, isn't it?
compare(Lemongrass Grill, Monsoon) Lemongrass Grill, which has poor atmosphere, quite bad atmosphere, is somewhat affordable. Monsoon is low-cost and it has like, nice atmosphere, it has nice atmosphere, you know.
compare(Lemongrass Grill, Monsoon) Basically, I mean, it seems that Lemongrass Grill is affordable, it's rather low-priced. Basically, I mean, Monsoon is low-cost. It doesn't feature rather nasty ambience!

compare(Lemongrass Grill, Monsoon) Monsoon's price is 26 dollars. It just has pleasant ambiance. It has acceptable ambience. Lemongrass Grill features kind of poor atmosphere, quite mediocre atmosphere, doesn't it?
compare(Lemongrass Grill, Monsoon) Ok, well, Monsoon features pleasant ambience, the ambiance is quite satisfactory. It seems that Lemongrass Grill has quite mediocre atmosphere.
compare(Lemongrass Grill, Monsoon) Right, Lemongrass Grill, which doesn't feature good ambience, is sort of quite low-priced. Monsoon is somewhat low-cost with acceptable ambiance, isn't it?
compare(Lemongrass Grill, Monsoon) Right, err... Lemongrass Grill is sort of low-cost. Monsoon's price is 26 dollars. Its price is 26 dollars. It doesn't have kind of nasty atmosphere.
compare(Les Routiers, Radio Perfecto) I see, yeah, Radio Perfecto, which has decent food, is darn low-cost. Basically, Les Routiers, on the other hand, offers quite adequate food. Great places.
compare(Les Routiers, Radio Perfecto) Did you say Les Routiers and Radio Perfecto? Oh gosh basically, I suppose they're really outstanding. Les Routiers is just located in Manhattan with decent food. However, Radio Perfecto's price is 25 dollars, the food is passable and it's in the East Village.
compare(Les Routiers, Radio Perfecto) Oh, although Les Routiers's price is around 41 dollars, the food is just acceptable. It seems that it's located in Manhattan. Radio Perfecto has like, satisfactory food, also its price is 25 dollars. I would consider them, they.
compare(Les Routiers, Radio Perfecto) Yeah, I thought everybody knew that Radio Perfecto is in the East Village, it's located in the East Village, the food is quite decent and its price is around 25 dollars while Les Routiers provides adequate food. They're darn great.
compare(Les Routiers, Radio Perfecto) Let's see what we can find on Les Routiers and Radio Perfecto. Basically, I guess you would like them. Radio Perfecto's price is 25 dollars and it has... the food is rather satisfactory. Les Routiers is mmhm... in Manhattan. Even if the food is somewhat satisfying, its price is 41 dollars.
compare(Les Routiers, Radio Perfecto) I am not sure. Actually, I know you would adore Les Routiers and Radio Perfecto! Radio Perfecto features like, decent food, sort of adequate food. Obviously, on the other hand, Les Routiers is located in Manhattan.
compare(Les Routiers, Radio Perfecto) I guess you would like Les Routiers and Radio Perfecto. Mmhm... Les Routiers offers like, really passable food. Basically, Radio Perfecto's price, on the other hand, is around 25 dollars and the food is adequate.
compare(Les Routiers, Radio Perfecto) Oh ok, I would recommend Les Routiers and Radio Perfecto. Err... Radio Perfecto is affordable, also it features... the food is acceptable. Les Routiers's price is... it's damn expensive.
compare(Mavalli Palace, Once Upon A Tart) Let's see what we can find on Mavalli Palace and Once Upon A Tart. Right, come on, Mavalli Palace is located in Murray Hill and an indian place, also its price is 27 dollars. Actually, Once Upon A Tart, however, is in TriBeCa/SoHo and it's damn cheap, you know. Acceptable restaurants.
compare(Mavalli Palace, Once Upon A Tart) Actually, I might be quite wrong. It seems that you would probably like Mavalli Palace and Once Upon A Tart. Obviously, Once Upon A Tart isn't somewhat expensive. However, Mavalli Palace's price is 27 dollars.
compare(Mavalli Palace, Once Upon A Tart) I don't know. I would suggest Mavalli Palace and Once Upon A Tart, you know. Mavalli Palace, which is an indian eating house and located in Murray Hill, is darn low-priced buddy. Once Upon A Tart is a cafe and sandwich restaurant. Mmhm... it's located in TriBeCa/SoHo.
compare(Mavalli Palace, Once Upon A Tart) Err... I am not sure. It seems that Mavalli Palace and Once Upon A Tart are like, fantastic places. I mean, Once Upon A Tart is somewhat cheap and it's in TriBeCa/SoHo while Mavalli Palace's price is 27 dollars and it's in Murray Hill, in Murray Hill.
compare(Mavalli Palace, Once Upon A Tart) Did you say Mavalli Palace and Once Upon A Tart? Mmhm... I suspect you would enjoy them. Obviously, err... Mavalli Palace is bloody low-cost. This restaurant, which is an indian place, is just located in Murray Hill. Once Upon A Tart is in TriBeCa/SoHo and a cafe and sandwich place.
compare(Mavalli Palace, Once Upon A Tart) Oh ok, Mavalli Palace's price is around 27 dollars while Once Upon A Tart is in TriBeCa/SoHo and it isn't expensive, this eating place is somewhat inexpensive. Err... I would approve them.
compare(Mavalli Palace, Once Upon A Tart) Let's see, Mavalli Palace and Once Upon A Tart... I see, ok, they're sa-sa-satisfactory restaurants. Ac-ac-actually, basically, Mavalli Palace's price is... its price is 27 dollars pal, but Once Upon A Tart is a cafe and sandwich eating house and located in TriBeCa/SoHo.
compare(Mavalli Palace, Once Upon A Tart) You want to know more about Once Upon A Tart and Mavalli Palace? Right, mmhm... Once Upon A Tart is a cafe and sandwich restaurant and located in TriBeCa/SoHo. Mavalli Palace is just located in Murray Hill, also it's... it's an indian place, and it's sort of low-cost, you know. They're outstanding.

compare(Old San Juan, Veselka) Err... actually, Veselka is located in Manhattan, you know but Old San Juan is in Manhattan.
compare(Old San Juan, Veselka) Oh God oh gosh Veselka, which is in Manhattan, is an east european eating house with quite rude staff. I mean, Old San Juan offers unmannered service, also it's a latin american place, okay? These eating places are like, satisfactory restaurants.
compare(Old San Juan, Veselka) Err... Old San Juan and Veselka are great. Actually, I suppose Old San Juan is a latin american place and in Manhattan, also it has rather bad staff. Veselka is an east european restaurant, also the waiters aren't somewhat good.
compare(Old San Juan, Veselka) You want to know more about Veselka and Old San Juan? Obviously, oh Veselka has unmannerly waiters and it's located in Manhattan, also it's an east european eating place. Mmhm... err... Old San Juan features sort of unmannered staff, also it's in Manhattan and a latin american restaurant.
compare(Old San Juan, Veselka) Come on, well, Veselka provides sort of unmannered service, doesn't it? Old San Juan, which features like, unmannerly service, ill-mannered waiters, you know, is a latin american restaurant and located in Manhattan.
compare(Old San Juan, Veselka) Old San Juan and Veselka are rather great. Err... Old San Juan is a latin american restaurant and in Manhattan, also the staff isn't really quite good. I mean, Veselka is an east european place with rude waiters.
compare(Old San Juan, Veselka) I suppose Old San Juan is in Manhattan with sort of unmannered waiters, kind of unmannerly waiters, you know. On the other hand, I mean, Veselka is located in Manhattan.
compare(Old San Juan, Veselka) You want to know more about Veselka and Old San Juan? I see, actually, I guess Veselka, which doesn't have quite good waiters, is in Manhattan, you see? Old San Juan, which is a latin american restaurant, is in Manhattan with sort of unmannerly staff. I would consider them.
compare(Scopa, Shabu, Tatsu) Let's see what we can find on Shabu-Tatsu and Scopa. Yeah, right, I thought everybody knew that Shabu-Tatsu has... the waiters are just nice, but this restaurant offers like, bad ambience, though. Scopa provides poor ambiance. I suspect they're alright.
compare(Scopa, Shabu, Tatsu) Let's see, Shabu-Tatsu and Scopa... Oh well, Shabu-Tatsu, which features like, acceptable waiters, is a japanese eating house while Scopa is an italian and new american eating place and the waiters aren't nasty, you know. Actually, you would probably adore them.
compare(Scopa, Shabu, Tatsu) Did you say Scopa and Shabu-Tatsu? Oh God ok, I would suggest them. Actually, Scopa offers really poor ambiance, although the waiters are somewhat friendly. Shabu-Tatsu, which provides mediocre ambience although this eating place has adequate waiters, is a japanese restaurant.
compare(Scopa, Shabu, Tatsu) Right, i see, you would probably enjoy Scopa and Shabu-Tatsu, you know. Scopa has kind of friendly staff, rather nice waiters. Shabu-Tatsu, which offers adequate service, however, is a japanese place.
compare(Scopa, Shabu, Tatsu) Right, oh gosh Shabu-Tatsu is a japanese eating house with welcoming staff, you know. Mmhm... actually, Scopa, which doesn't provide good ambiance even if this eating place features pa-pa-passable waiters, is an italian and new american restaurant.
compare(Scopa, Shabu, Tatsu) Right, everybody knows that you would probably adore Scopa and Shabu-Tatsu. Although Scopa features poor atmosphere, basically, it has adequate waiters, nice waiters. Err... this restaurant is an italian and new american place. Shabu-Tatsu provides friendly service, although the ambience isn't somewhat good.
compare(Scopa, Shabu, Tatsu) Let's see, Shabu-Tatsu and Scopa... Yeah, ok, Shabu-Tatsu features satisfactory service. However, Scopa doesn't provide sort of nasty staff. Actually, basically, I would recommend them.
compare(Scopa, Shabu, Tatsu) Well, yeah, I would consider Scopa and Shabu-Tatsu. It seems to me that Shabu-Tatsu features acceptable waiters. However, Scopa just has really mediocre ambience.
compare(Soul Fixins', Walker's) I see, come on, Soul Fixins' is a southern eating place with unmannerly staff, ill-mannered waiters. I mean, it seems that Walker's is a bar snacks restaurant. Even if it's damn low-priced, this eating house has unmannered waiters.
compare(Soul Fixins', Walker's) Did you say Walker's and Soul Fixins'? I might be darn wrong. Basically, Walker's's price is around 25 dollars. Soul Fixins''s price, on the other hand, is 15 dollars! Mmhm... I might suggest them.
compare(Soul Fixins', Walker's) Ok, well, I would consider Soul Fixins' and Walker's, I might su-suggest them, mate. Even if Walker's doesn't have good waiters, it's really affordable. This eating house is a bar snacks eating place. It seems to me that So-So-Soul Fixins' is a southern restaurant and the service is unmannerly.

compare(Soul Fixins', Walker's) Let's see, Walker's and Soul Fixins'... I don't know! Actually, Walker's is somewhat affordable while Soul Fixins''s price is... it's quite inexpensive. Mmhm... I might consider these eating houses.
compare(Soul Fixins', Walker's) Mmhm... I am not quite sure. Err... Walker's has bad service, bad waiters, even if it's kind of low-cost. It seems that Soul Fixins' has bad waiters.
compare(Soul Fixins', Walker's) Acceptable places. It seems that Walker's, which fe-features rather bad service, kind of bad waiters, even if this restaurant's price is around 25 dollars, is a bar snacks place. Err... Soul Fixins' doesn't have good waiters.
compare(Soul Fixins', Walker's) I see, Walker's, which is affordable, is a bar snacks place but Soul Fixins' is a southern restaurant, also it's somewhat cheap. Actually, basically, I would recommend them, you know.
compare(Soul Fixins', Walker's) I am not sure, would you be? Basically, it seems to me that Soul Fixins', which features really kind of unmannerly waiters, is a southern restaurant. Even if Walker's provides rather unmannered service, its price is 25 dollars. I believe you would love them.
recommend(Amy's Bread) Err... actually, Amy's Bread is one of my favourite places, one of my favourite places, with good food, good food, isn't it?
recommend(Amy's Bread) I mean, it seems to me that Amy's Bread has like, bad atmosphere, and nice servers. Actually, it's the only place I would recommend! It isn't as bad as the others, is it?
recommend(Amy's Bread) Amy's Bread just has nice servers and good food. It's cheap, although the atmosphere is bad, its price is 12 dollars. It isn't as bad as the others.
recommend(Amy's Bread) I am sort of quite sure you would like Amy's Bread.
recommend(Amy's Bread) Amy's Bread is one of my favourite restaurants. The food isn't kind of nasty. It features adequate service. Err... the staff is acceptable.
recommend(Amy's Bread) Amy's Bread isn't as bad as the others. It seems that it's cheap. The food is good and it has friendly staff, acceptable service!
recommend(Amy's Bread) I think that Amy's Bread isn't as bad as the others, it's the only restaurant that is any good.
recommend(Amy's Bread) Right, I think that Amy's Bread, which doesn't feature kind of quite bad food, is in Midtown West. Basically, it's the only restaurant I would recommend.
recommend(Amy's Bread) Ok, I mean, Amy's Bread, which is really inexpensive with tasty food, really nice food, is a kind of cafes eating place with kind of friendly staff. I mean, it's the only restaurant I would recommend.
recommend(Amy's Bread) Ok, Amy's Bread features satisfactory waiters, doesn't it? It has really nice food, the food isn't really nasty, and its price is 12 dollars. It's the only restaurant I would recommend.
recommend(Amy's Bread) Basically, Amy's Bread, which is the only place I would recommend, is just quite located in Midtown West, in Midtown West, with friendly servers, nice servers.
recommend(Amy's Bread) I am sure you would like Amy's Bread. Its price is around 12 dollars and the food is good. This restaurant, which has nice servers, is a cafes place.
recommend(Azuri Cafe) Azuri Cafe is the best restaurant of its kind. Its price is 14 dollars, it's really cheap. It features good food, it has nice food. It's a kosher and vegetarian place, isn't it?
recommend(Azuri Cafe) Azuri Cafe isn't as bad as the others. It's the only restaurant I would recommend. It's rather kind of inexpensive.
recommend(Azuri Cafe) Azuri Cafe features rude waiters. The food is nice, although it doesn't have good ambience. It's the only restaurant I would recommend.

recommend(Azuri Cafe) Yeah, Azuri Cafe features rude waiters, it's a kosher and vegetarian restaurant and the food is nice. Its price is 14 dollars, even if the ambience is poor. I am sure you would like it, it's the best.
recommend(Azuri Cafe) I see, err... Azuri Cafe is one of my favourite restaurants, it's one of my favourite places. This restaurant, which doesn't have good atmosphere even if its price is 14 dollars, is like, a really kosher and vegetarian place, with good food and bad servers.
recommend(Azuri Cafe) Yeah, basically, I think that Azuri Cafe doesn't feature sort of kind of bad food. Basically, it's above the competition, the sort of kind of best restaurant, isn't it?
recommend(Bond Street) Well, Bond Street, which has great food, great food, even if it's kind of expensive, is a japanese and sushi place. Actually, it's the only place that is any good, it's the really best place.
recommend(Bond Street) Bond Street doesn't feature quite nasty staff, does it? I think that it isn't as bad as the others, I am quite sure you would like it.
recommend(Bond Street) I see, Bond Street is a japanese and sushi restaurant, isn't it? Its price is 51 dollars. It's costly. It has excellent food. It's one of my favourite places.
recommend(Bond Street) Well, ok, Bond Street is a japanese and sushi place. It isn't quite cheap! However, the food is excellent. It isn't as bad as the others, is it? Basically, it's the only restaurant that is any good.
recommend(Bond Street) Bond Street's price is 51 dollars and it doesn't feature bad ambience. It has wonderful food. It has excellent food. Actually, it's the only restaurant I would recommend.
recommend(Bond Street) Right, Bond Street is the best place of its kind, you know. It has kind of excellent food, kind of great food, you know!
recommend(Cent anni) Did you say Ce-Cent'anni? I see, I mean, I would consider it because it has friendly staff and tasty food, you know buddy.
recommend(Cent anni) I am not sure. Cent'anni is... it's located in Manhattan, also the atmosphere is somewhat bloody poor, but it features tasty food though. Actually, this eating house, which provides quite acceptable service, is an italian restaurant. It's sort of the best eating place of its kind.
recommend(Cent anni) I don't know! Err... I mean, I suppose Cent'anni is one of my favourite places, alright?
recommend(Cent anni) Because Cent'anni is an italian restaurant with quite decent staff and rather nice food, I mean, I believe it's darn alright.
recommend(Cent anni) Yeah, with sort of good food, Cent'anni is quite the best place of its kind pal!
recommend(Cent anni) I see, even if Cent'anni is somewhat expensive, it provides tasty food. It seems that this eating place, which offers quite friendly service even if the atmosphere is damn poor, is located in Manhattan. I would advise this restaurant.
recommend(Cent anni) Err... I don't know! Because Cent'anni features quite nice waiters, I would advise it.
recommend(Cent anni) I mean, I might be sort of wrong. Because Cent'anni provides like, quite nice service, friendly waiters, I imagine you would appreciate it, you know.
recommend(Chanpen Thai) Right, I mean, Chanpen Thai has nice food, the food is adequate. This restaurant, which has friendly servers, is low-cost. I mean, it's one of my favourite places.
recommend(Chanpen Thai) Yeah, Chanpen Thai is a quite thai place with kind of nice atmosphere. Its price is 24 dollars and the food is adequate with nice servers. It's the best, isn't it?
recommend(Chanpen Thai) I see, actually, Chanpen Thai is just rather low-cost with really acceptable food. Actually, it has rather friendly staff, it features really adequate service. It isn't as bad as the others, it's the best.

recommend(Chanpen Thai) It seems to me that I am sure you would like Chanpen Thai. It features passable food, satisfying food, doesn't it?
recommend(Chanpen Thai) I am sure you would like Chanpen Thai. I mean, the staff is adequate, it doesn't feature really nasty waiters, and the food is rather satisfactory with acceptable ambience.
recommend(Chanpen Thai) Chanpen Thai is a thai restaurant and its price is 24 dollars. Basically, it's the only place I would recommend.
recommend(Chimichurri Grill) Obviously, oh God Chimichurri Grill has like, quite adequate food, so it's the only re-restaurant I would consider, and it's a latin american place and located in Midtown West with ra-ra-rather friendly staff.
recommend(Chimichurri Grill) Oh God well, Chimichurri Grill has mediocre ambiance, the waiters are really sort of passable though and it's a latin american restaurant! Even if the food is adequate, this eating house's price is 41 dollars. You would probably enjoy it.
recommend(Chimichurri Grill) I would suggest... because this restaurant has adequate food, acceptable food, and rather friendly service, I guess you would like it mate, you know!
recommend(Chimichurri Grill) I see, oh Chimichurri Grill is a latin american place with sort of poor atmosphere. Although it doesn't have rather nasty food, its price is 41 dollars. I suspect it's kind of alright.
recommend(Chimichurri Grill) I see, because Chimichurri Grill is a latin american eating house, I would approve it, okay?
recommend(Chimichurri Grill) Oh err... I think that Chimichurri Grill is the only restaurant that is any good, you know buddy.
recommend(Chimichurri Grill) Err... I might be rather wrong. Because Chimichurri Grill is quite pricy, but the food is somewhat decent though, also this restaurant is a latin american place with passable waiters, adequate staff, and mediocre ambiance, you see? basically, it's the only place that is any good.
recommend(Chimichurri Grill) Basically, Chimichurri Grill is err... in Midtown West, so I guess it's really kind of alright, it's a latin american restaurant, it offers quite mediocre ambience and its price is 41 dollars.
recommend(Edgar's Cafe) Did you say Edgar's Cafe? Mmhm... basically, I might be wrong. Because Edgar's Cafe, which doesn't offer nasty food, is darn inexpensive, I guess you would like this eating house, it's really alright!
recommend(Edgar's Cafe) Err... oh I thought everybody knew that Edgar's Cafe offers ill-mannered service, so I guess you would like it, buddy and it's located in Manhattan with quite second-rate atmosphere, also it's a cafe eating place.
recommend(Edgar's Cafe) Err... I suppose Edgar's Cafe offers kind of acceptable food, so it isn't as bad as the others, and its price is... this eating house's price is around 19 dollars, you know.
recommend(Edgar's Cafe) Let's see what we can find on Edgar's Cafe. Mmhm... I am not rather sure pal. I mean, basically, Edgar's Cafe isn't as bad as the others.
recommend(Edgar's Cafe) Ob-ob-obviously, ba-ba-basically, Edgar's Cafe's price is 19 dollars, so I know this restaurant is sort of alright, and it's in Manhattan and a cafe place with nice food mate, you know.
recommend(Edgar's Cafe) Well, I would advise Edgar's Cafe. It offers kind of poor ambience. I think that this restaurant provides nice food, even if the service is somewhat damn bad.
recommend(Edgar's Cafe) Err... actually, I don't know. Edgar's Cafe, which is located in Manhattan with sort of rude waiters and rather bad ambience, is the only place that is any good!
recommend(Edgar's Cafe) Oh gosh ok, Edgar's Cafe is the best restaurant of its kind since this eating place, which doesn't offer good service, is in Manhattan with quite mediocre atmosphere, you know!
recommend(Flor De Mayo) Flor De Mayo isn't as bad as the others, is it? Its price is 18 dollars and it features nice food with kind of decent servers, satisfactory waiters.
recommend(Flor De Mayo) Right, i see, Flor De Mayo is one of my favourite eating places. It's the only restaurant I would recommend, you know. I think that this restaurant, which features nice food with friendly staff, is somewhat inexpensive, you know.

recommend(Flor De Mayo) Flor De Mayo is the only eating house that is any good. I mean, it's inexpensive. I mean, it features nice food, the servers are friendly, the waiters are acceptable, it's located in Uptown Manhattan and it's a rather chinese and latin american eating place, it's a rather chinese and latin american restaurant.
recommend(Flor De Mayo) I see, basically, Flor De Mayo is like, the rather best restaurant, above the competition, with sort of decent waiters, satisfactory staff, isn't it?
recommend(Flor De Mayo) It seems to me that Flor De Mayo is the only eating house I would recommend. This restaurant, which doesn't feature nasty food, is inexpensive. The servers are quite friendly, the waiters are adequate, it's a chinese and latin american eating place, it's a chinese and latin american restaurant, it's quite located in Uptown Manhattan and the ambience is mediocre.
recommend(Flor De Mayo) Err... I mean, I am quite sure you would like Flor De Mayo. It's cheap, it's just cheap, and the food is good. It seems that the service isn't nasty. It's a chinese and latin american place and in Uptown Manhattan.
recommend(Flor De Mayo) Let's see what we can find on Flor De Mayo. I am not quite sure. I would approve Flor De Mayo, I would recommend it. I mean, actually, its price is around 18 dollars, even if this eating house doesn't feature sort of good ambience.
recommend(Flor De Mayo) Ok, basically, Flor De Mayo offers poor ambience, so I believe it's alright, but it's kind of cheap though, also the staff is rather bad, it provides nice food though and this restaurant is in the Upper West Side, you know.
recommend(Flor De Mayo) Let's see what we can find on Flor De Mayo. Obviously, ok, it's a chinese and latin american place, so it's the only place I would consider, also it has... the food isn't quite nasty, and its price is 18 dollars, its price is 18 dollars!
recommend(Flor De Mayo) Did you say Flor De Mayo? Yeah, well, it's the best place of its kind. Actually, the atmosphere is somewhat second-rate, although this eating house is darn inexpensive. This restaurant, which has ill-mannered waiters, is a chinese and latin american eating place and in the Upper West Side.
recommend(Flor De Mayo) Mmhm... actually, I don't know. Flor De Mayo provides acceptable food, so I believe you would love this eating place, I imagine it's quite alright, pal, and its price is 18 dollars.
recommend(Flor De Mayo) Did you say Flor De Mayo? I am not sure. Basically, I guess Flor De Mayo is darn alright, I would suggest it. Although this eating house has really kind of mediocre ambience, it's quite inexpensive.
recommend(Flor De Mayo) You would probably appreciate Flor De Mayo. Mmhm... the staff is kind of bad, although the food is somewhat nice. I think that this restaurant, which offers poor ambience even if its price is 18 dollars, is located in the Upper West Side!
recommend(Flor De Mayo) Basically, it seems that Flor De Mayo is the only restaurant that is any good since it's a chinese and latin american eating place, you know.
recommend(Jin Dal Lae) Err... actually, I don't know. Because Jin Dal Lae is damn low-cost with kind of nice food, basically, I guess you would like it, okay?
recommend(Jin Dal Lae) You want to know more about Jin Dal Lae? Ok, it has quite nice food, even if it offers really poor ambience. I guess it's a korean place, its price is around 27 dollars, but the staff is bad though. I suspect it's alright.
recommend(Jin Dal Lae) Did you say Jin Dal Lae? Come on, i see, it has quite poor ambience, so I suspect you would enjoy this restaurant, also it's a korean place, a korean place, you know!
recommend(Jin Dal Lae) You want to know more about Jin Dal Lae? Oh because Jin Dal Lae offers quite second-rate atmosphere, but the food is sort of acceptable though, also it's... it's a korean restaurant, you know, alright? it's one of my favourite eating houses!
recommend(Jin Dal Lae) I suspect Jin Dal Lae is alright, it's the only place I would consider, you see? Although it features kind of bad waiters, it's low-cost. It seems that this restaurant is a korean place. Mmhm... although it just has bad atmosphere, the food is... the food is quite nice.
recommend(Jin Dal Lae) It seems that Jin Dal Lae is... its price is 27 dollars mate. This eating place, which provides like, sort of acceptable food, is in The Bronx. I would approve it, wouldn't you?
recommend(Jin Dal Lae) Basically, I would approve... since its price is 27 dollars pal, it isn't as bad as the others.

Appendix C

Utterances Generated using PERSONAGE-PE

This appendix contains the 50 utterances that were generated using PERSONAGE-PE for the evaluation presented in Chapter 9. Each utterance was generated with extreme values for two traits, and mid-range values for the remaining three traits (on a scale from 1 to 7).

Content plan	Trait values
compare(Caffe Cielo, Trattoria Spagh.)	**extra=3.72, ems=4.56, agree=1.00, consc=1.00, open=3.40**
You would probably appreciate Caffe Cielo and Trattoria Spaghetto. Caffe Cielo is an italian eating place. It offers quite adequate food, even if it provides kind of mediocre ambience. Trattoria Spaghetto is an italian restaurant. Err... the ambiance is damn poor, isn't it?	
compare(Caffe Cielo, Trattoria Spagh.)	**extra=3.98, ems=3.42, agree=3.56, consc=6.65, open=1.71**
Caffe Cielo and Trattoria Spaghetto are alright. Caffe Cielo offers mediocre ambiance, even if it provides adequate food. Trattoria Spaghetto is an italian restaurant with poor ambience.	
compare(Caffe Cielo, Trattoria Spagh.)	**extra=5.83, ems=6.02, agree=7.00, consc=4.47, open=3.86**
Caffe Cielo offers acceptable food. Trattoria Spaghetto provides adequate food mate.	
compare(Caffe Cielo, Trattoria Spagh.)	**extra=7.00, ems=3.96, agree=4.05, consc=4.76, open=6.89**
Although Trattoria Spaghetto offers acceptable food, it features poor ambiance! This eating house is an italian restaurant. Even if Caffe Cielo provides adequate food, the ambience is mediocre.	
compare(Caffe Cielo, Trattoria Spagh.)	**extra=7.00, ems=7.00, agree=4.11, consc=3.91, open=3.08**
Even if Trattoria Spaghetto offers satisfying food, the ambiance is mediocre. Caffe Cielo provides kind of second-rate ambience, even if it has acceptable food. It's an italian restaurant!	
compare(Les Routiers, Radio Perfecto)	**extra=3.07, ems=3.88, agree=1.28, consc=4.43, open=7.00**
Les Routiers and Radio Perfecto are outstanding. Radio Perfecto's price is 25 dollars, isn't it? The food is acceptable, also it's located in the East Village. Err... although Les Routiers offers adequate food, it's damn costly.	
compare(Les Routiers, Radio Perfecto)	**extra=4.53, ems=6.88, agree=2.95, consc=5.59, open=6.52**
Outstanding restaurants. Radio Perfecto offers acceptable food, also it's in the East Village. Les Routiers is in Manhattan, also its price is 41 dollars.	
compare(Les Routiers, Radio Perfecto)	**extra=4.89, ems=7.00, agree=3.55, consc=6.21, open=2.88**
Let's see, Les Routiers and Radio Perfecto... You would probably appreciate them. Radio Perfecto is in the East Village with kind of acceptable food. Les Routiers is located in Manhattan. Its price is 41 dollars.	
compare(Les Routiers, Radio Perfecto)	**extra=6.90, ems=4.14, agree=7.00, consc=3.10, open=3.08**
Radio Perfecto's price is 25 dollars but Les Routiers provides adequate food. I imagine they're alright!	
compare(Les Routiers, Radio Perfecto)	**extra=7.00, ems=5.44, agree=4.20, consc=6.78, open=3.72**
Even if Les Routiers provides adequate food, it's costly. Radio Perfecto offers kind of acceptable food and its price is 25 dollars. I would suggest them!	
compare(Old San Juan, Veselka)	**extra=3.32, ems=1.00, agree=4.24, consc=4.10, open=7.00**
Veselka has bad wa-waiters, also it's in Manhattan. Err... Old San Juan provides rude staff, also it's located in Midtown West. I would suggest them.	
compare(Old San Juan, Veselka)	**extra=4.30, ems=1.00, agree=3.56, consc=1.20, open=3.44**
Ok, you would probably appreciate Old San Juan and Veselka. Old San Juan is err... in Midtown West, also it provides quite rude staff, you see? Veselka is in Manhattan, also it has bad waiters.	
compare(Old San Juan, Veselka)	**extra=4.91, ems=4.38, agree=6.30, consc=3.69, open=1.00**
Veselka is located in Manhattan. It's an east european restaurant. Old San Juan is located in Midtown West. They're the only eating houses I would suggest.	
compare(Old San Juan, Veselka)	**extra=7.00, ems=4.25, agree=1.43, consc=4.23, open=3.82**
Old San Juan is in Midtown West. It's a latin american eating house, also this eating place provides unmannerly staff! Veselka is an east european restaurant, also the waiters are bloody unmannered, okay?	
compare(Old San Juan, Veselka)	**extra=7.00, ems=4.78, agree=4.62, consc=7.00, open=4.85**
Veselka offers kind of unmannerly staff! It's located in Manhattan, also this eating house is an east european restaurant. Old San Juan features unmannered waiters. I would suggest them.	
compare(Scopa, Shabu-Tatsu)	**extra=3.09, ems=4.58, agree=6.93, consc=7.00, open=3.73**
Shabu-Tatsu offers friendly staff but Scopa has adequate waiters. I might suggest them.	
compare(Scopa, Shabu-Tatsu)	**extra=3.48, ems=6.25, agree=2.13, consc=5.05, open=4.18**
Outstanding eating places. Scopa is an italian and new american place with second-rate ambiance. Err... Shabu-Tatsu is a japanese restaurant. It offers mediocre ambience, even if this eating house features adequate waiters.	
compare(Scopa, Shabu-Tatsu)	**extra=4.05, ems=4.34, agree=3.08, consc=1.09, open=1.35**
Scopa and Shabu-Tatsu are quite alright. Shabu-Tatsu is a japanese restaurant, isn't it? Err... the waiters are friendly, even if it provides poor ambience. Scopa offers mediocre ambiance.	
compare(Scopa, Shabu-Tatsu)	**extra=6.14, ems=4.41, agree=5.16, consc=3.45, open=6.20**
You want to know more about Shabu-Tatsu and Scopa? Even if Shabu-Tatsu features mediocre ambiance, it offers friendly staff. It's a japanese restaurant. Even if Scopa provides poor ambience, the waiters are adequate. I imagine you would appreciate these eating houses.	
compare(Scopa, Shabu-Tatsu)	**extra=6.90, ems=1.25, agree=4.29, consc=5.15, open=4.11**
Ok, even if Shabu-Tatsu provides bad ambiance, it has friendly service. Scopa offers nice staff, although the ambience is damn poor. I would suggest them!	

<table>
<tr><th>Content plan</th><th>Trait values</th></tr>
<tr><td>compare(Soul Fixins', Walker's)</td><td>extra=3.32, ems=4.80, agree=5.36, consc=7.00, open=1.00</td></tr>
<tr><td colspan="2">Walker's' price is 25 dollars. Soul Fixins' is a southern restaurant and it's inexpensive. I might suggest these eating houses.</td></tr>
<tr><td>compare(Soul Fixins', Walker's)</td><td>extra=4.37, ems=4.29, agree=7.00, consc=7.00, open=4.08</td></tr>
<tr><td colspan="2">Walker's' price is 25 dollars. Soul Fixins' is inexpensive. I imagine they're alright.</td></tr>
<tr><td>compare(Soul Fixins', Walker's)</td><td>extra=4.86, ems=6.67, agree=6.84, consc=4.62, open=4.07</td></tr>
<tr><td colspan="2">Walker's' price is 25 dollars. Soul Fixins' is inexpensive pal. I would suggest them.</td></tr>
<tr><td>compare(Soul Fixins', Walker's)</td><td>extra=7.00, ems=1.00, agree=4.79, consc=4.35, open=3.65</td></tr>
<tr><td colspan="2">Did you say Walker's and Soul Fixins'? Ok, Walker's is a bar snacks place, also its price is 25 dollars. Soul Fixins' is a southern place! Its price is... this restaurant is bloody cheap. I would suggest them, wouldn't you?</td></tr>
<tr><td>compare(Soul Fixins', Walker's)</td><td>extra=7.00, ems=4.53, agree=3.38, consc=4.48, open=1.84</td></tr>
<tr><td colspan="2">Walker's' price is 25 dollars. It provides unmannered staff. This eating house is a bar snacks restaurant. Soul Fixins' is a southern eating place with unmannerly waiters!</td></tr>
<tr><td>recommend(Chimichurri Grill)</td><td>extra=1.00, ems=4.44, agree=3.34, consc=4.10, open=1.12</td></tr>
<tr><td colspan="2">Err... Chimichurri Grill is the only restaurant that is any good. Even if this eating place is costly, it offers adequate food, alright? It provides mediocre ambience.</td></tr>
<tr><td>recommend(Chimichurri Grill)</td><td>extra=4.10, ems=6.85, agree=7.00, consc=3.58, open=4.34</td></tr>
<tr><td colspan="2">I would suggest Chimichurri Grill.</td></tr>
<tr><td>recommend(Chimichurri Grill)</td><td>extra=4.18, ems=3.38, agree=1.00, consc=1.97, open=4.59</td></tr>
<tr><td colspan="2">You would probably appreciate Chimichurri Grill. Err... it's a latin american restaurant, also the ambience is damn poor. It offers friendly staff, even if it's quite costly alright?</td></tr>
<tr><td>recommend(Chimichurri Grill)</td><td>extra=5.00, ems=2.92, agree=4.48, consc=1.00, open=7.00</td></tr>
<tr><td colspan="2">Chimichurri Grill is a latin american restaurant, also it's located in Midtown West. It has quite friendly waiters. It offers adequate food. I imagine you would appreciate it.</td></tr>
<tr><td>recommend(Chimichurri Grill)</td><td>extra=6.56, ems=1.51, agree=4.00, consc=4.89, open=4.03</td></tr>
<tr><td colspan="2">Ok, Chimichurri Grill offers poor ambience, even if it provides... it has nice waiters! Although the food is adequate, it's damn costly. This restaurant is one of my favourite places.</td></tr>
<tr><td>recommend(Jin Dal Lae)</td><td>extra=3.34, ems=7.00, agree=4.07, consc=1.60, open=3.99</td></tr>
<tr><td colspan="2">Jin Dal Lae offers quite acceptable food, even if it provides mediocre ambience. It's a korean restaurant. You would probably appreciate it.</td></tr>
<tr><td>recommend(Jin Dal Lae)</td><td>extra=3.75, ems=6.32, agree=3.91, consc=3.95, open=1.00</td></tr>
<tr><td colspan="2">Jin Dal Lae isn't as bad as the others. It offers mediocre ambience, even if this eating place's price is 27 dollars. It features unmannerly waiters.</td></tr>
<tr><td>recommend(Jin Dal Lae)</td><td>extra=4.34, ems=4.26, agree=7.00, consc=4.47, open=1.55</td></tr>
<tr><td colspan="2">Jin Dal Lae is the only restaurant I would suggest.</td></tr>
<tr><td>recommend(Jin Dal Lae)</td><td>extra=5.69, ems=3.98, agree=4.43, consc=1.25, open=2.63</td></tr>
<tr><td colspan="2">Ok, Jin Dal Lae is a korean restaurant. It's... its price is 27 dollars, also it's in The Bronx with kind of adequate food. I imagine this eating house is damn alright, you see?</td></tr>
<tr><td>recommend(Jin Dal Lae)</td><td>extra=6.67, ems=4.64, agree=1.02, consc=4.06, open=3.96</td></tr>
<tr><td colspan="2">Oh God err... Jin Dal Lae is the only restaurant that is any good, alright? Even if this eating place provides adequate food, it offers mediocre ambience! This restaurant is a korean eating house, also the waiters are unmannerly.</td></tr>
<tr><td>recommend(Pepolino)</td><td>extra=1.00, ems=6.49, agree=3.76, consc=3.62, open=2.81</td></tr>
<tr><td colspan="2">Pepolino isn't as bad as the others. Err... it features acceptable waiters, even if this eating house's price is 39 dollars. It offers mediocre ambience.</td></tr>
<tr><td>recommend(Pepolino)</td><td>extra=3.22, ems=1.19, agree=7.00, consc=3.68, open=3.93</td></tr>
<tr><td colspan="2">Ok, err... Pepolino provides good food, so I would suggest it, okay?</td></tr>
<tr><td>recommend(Pepolino)</td><td>extra=4.29, ems=4.53, agree=2.92, consc=1.12, open=1.11</td></tr>
<tr><td colspan="2">Ok, Pepolino is the only eating place that is any good. It's an italian restaurant. Err... it's located in TriBeCa/SoHo, isn't it? It's bloody costly with quite mediocre ambience.</td></tr>
<tr><td>recommend(Pepolino)</td><td>extra=4.78, ems=4.16, agree=1.00, consc=7.00, open=3.26</td></tr>
<tr><td colspan="2">Oh God Pepolino is the only restaurant that is any good, alright? This eating place is an italian eating house and in TriBeCa/SoHo with poor ambience. Err... its price is 39 dollars.</td></tr>
<tr><td>recommend(Pepolino)</td><td>extra=6.23, ems=3.51, agree=3.58, consc=2.75, open=1.13</td></tr>
<tr><td colspan="2">I imagine Pepolino is kind of alright. It offers friendly staff, also it's an italian place, alright? This restaurant's price is 39 dollars, although it provides quite tasty food.</td></tr>
<tr><td>recommend(Trattoria Rustica)</td><td>extra=1.46, ems=5.04, agree=3.21, consc=1.00, open=4.18</td></tr>
<tr><td colspan="2">There could be worse restaurants. Although Trattoria Rustica's price is 35 dollars, it has kind of friendly waiters. Err... it offers quite mediocre ambience, doesn't it?</td></tr>
<tr><td>recommend(Trattoria Rustica)</td><td>extra=1.92, ems=3.40, agree=1.73, consc=3.50, open=4.20</td></tr>
<tr><td colspan="2">Err... you would probably appreciate Trattoria Rustica, wouldn't you? It's in Manhattan, also it's an italian restaurant. It offers poor ambience, also it's quite costly.</td></tr>
</table>

Content plan	Trait values
recommend(Trattoria Rustica)	**extra=2.94, ems=1.97, agree=3.44, consc=3.41, open=1.00**
Trattoria Rustica isn't as bad as the others. Err... even if it's costly, it offers kind of adequate food, alright? It's an italian place.	
recommend(Trattoria Rustica)	**extra=3.94, ems=3.45, agree=1.00, consc=4.12, open=1.33**
Trattoria Rustica is the only eating place that is any good. Err... it's located in Manhattan. This restaurant is an italian place with poor ambience. It's bloody costly, even if this eating house has friendly waiters you see?	
recommend(Trattoria Rustica)	**extra=4.46, ems=1.00, agree=3.76, consc=6.73, open=2.73**
Let's see, Trattoria Rustica... Oh God err... you would probably appreciate it. It's costly, even if this restaurant offers adequate food. Even if it features nice waiters, the ambience is poor.	
recommend(Vinnie's Pizza)	**extra=1.00, ems=1.82, agree=3.78, consc=4.59, open=3.45**
Vinnie's Pizza isn't as bad as the others, is it? It's located in Manhattan. Err... even if this restaurant features kind of bad waiters, it offers adequate food.	
recommend(Vinnie's Pizza)	**extra=3.75, ems=3.59, agree=6.95, consc=6.64, open=3.96**
Yeah, Vinnie's Pizza is cheap, so I imagine it's alright.	
recommend(Vinnie's Pizza)	**extra=4.20, ems=4.84, agree=3.92, consc=7.00, open=5.48**
You would probably appreciate Vinnie's Pizza. It's a pizza restaurant and in Manhattan, also this eating place offers unmannered staff with terrible ambience.	
recommend(Vinnie's Pizza)	**extra=6.01, ems=6.22, agree=1.00, consc=3.37, open=3.28**
Oh God you would probably appreciate Vinnie's Pizza. It's a pizza restaurant, isn't it? This eating place is err... in Manhattan, also it has quite unmannerly waiters with terrible ambience.	
recommend(Vinnie's Pizza)	**extra=7.00, ems=4.10, agree=4.12, consc=4.06, open=7.00**
Vinnie's Pizza is a pizza restaurant, also it provides adequate food! Although its price is 13 dollars, it features kind of unmannered waiters. I imagine you would appreciate it.	

Bibliography

G. W. Allport and H. S. Odbert. Trait names: a psycho-lexical study. *Psychological Monographs*, 47(1, Whole No. 211):171–220, 1936. 2.1.1, 3.7, 6.4

E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. The automated design of believable dialogues for animated presentation teams. In J. S. S. Prevost J. Cassell and E. Churchill, editors, *Embodied conversational agents*, pages 220–255. MIT Press, Cambridge, MA, 2000. 2.4.3.4

S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker. Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005. 2.3.3, 3.2.4, 3.6

G. Ball and J. Breese. Emotion and personality in a conversational character. In *Proceedings of the Workshop on Embodied Conversational Characters*, pages 83–86, 1998. 2.4.3.4

S. Bangalore and O. Rambow. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 42–48, 2000. 8.1, 8.5

J. A. Bateman. KPML: The KOMET-Penman multilingual resource development environment. In *Proceedings of the 5th European Workshop on Natural Language Generation*, pages 219–222, 1995. 2.4

K. Beaman. Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D. Tannen and R. Freedle, editors, *Coherence in Spoken and Written Discourse*, pages 45–80. Ablex, 1984. 5.5.2

A. Bell. Language style as audience design. In N. Coupland and A. Jaworski, editors, *Sociolinguistics: A reader*, pages 240–250. St. Martin's Press, 1997. 1.3.1

A. Belz. Corpus-driven generation of weather forecasts. In *Proceedings of the 3rd Corpus Linguistics Conference*, 2005a. 2.4.4.2

A. Belz. Statistical generation: Three methods compared and evaluated. In *Proceedings of the 10th European Workshop on Natural Language Generation*, 2005b. 2.4.4.2

D. Biber. *Variation across Speech and Writing*. Cambridge University Press, 1988. 2.2.1

P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5 (9/10):341–345, 2001. 3.2.2.3

T. J. Bouchard and M. McGue. Genetic and environmental influences on human psychological differences. *Journal of Neurobiology*, 54:445, 2003. 2.1.2

H. P. Branigan, M. J. Pickering, and A. A. Cleland. Syntactic coordination in dialogue. *Cognition*, 75:B13–B25, 2000. 3.1

S. E. Brennan. Lexical entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialogue*, pages 41–44, 1996. 3.1, 5.5.5

P. Brown and S. Levinson. *Politeness: Some universals in language usage*. Cambridge University Press, 1987. 1.3.2.2, 2.4.3.3, 4.3, 4.5, 5.5.4, 7.2.1

E. Brunswik. *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley, CA, 1956. 4.1.2

D. Byrne and D. Nelson. Attraction as a linear function of proportion of positive reinforcements. *Journal of Personality and Social Psychology*, 1:659–663, 1965. 1.4, 3.1, 3.2.2.3

E. Campana, M. K. Tanenhaus, J. F. Allen, and R. W. Remington. Evaluating cognitive load in spoken language interfaces using a dual-task paradigm. In *Proceed-*

ings of the 9th International Conference on Spoken Language Processing (ICSLP), 2004. 1.4.2

G. Carenini and J. D. Moore. A strategy for generating evaluative arguments. In *Proceedings of International Conference on Natural Language Generation*, pages 47–54, Mitzpe Ramon, Israel, 2000. 4.3, 5.5.1

J. Cassell and T. Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13:89–132, 2003. 2.4.3.4, 3, 9.3

T. Chklovski and P. Pantel. VERBOCEAN: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, 2004. 5.5.5

J. Chu-Carroll. MIMIC: An adaptive mixed initiative spoken dialogue system for information queries. In *Proceedings of the Conference on Applied Natural Language Processing (ANLP)*, pages 97–104, 2000. 2.3.2

M. H. Cohen, J. P. Giangola, and J. Baloch. *Voice user interface design*, chapter 6, pages 75–83. Addison Wesley, 2004. 1.4.2

K. M. Colby. *Artificial paranoia: a computer simulation of paranoid processes*. Pergamon Press, New York, 1975. 2.4.1

M. Coltheart. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33A:497–505, 1981. 3.4, 3.2.2.1, 3.2.2.1, 7.2.2

C. Cope. Linguistic structure and personality development. *Journal of Counselling Psychology*, 16:1–19, 1969. 4.1, 4.3

P. T. Costa and R. R. McCrae. *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, FL, 1992. 3.1, 6.2, 6.4

J. A. Daly and A. M. Bippus. Personality and interpersonal communication. In J. C. McCroksey, J. A. Daly, M. M. Martin, and M. J. Beatty, editors, *Communication* *and Personality: Trait Perspectives*, chapter 1, pages 1–40. Hampton Press, 1998. 7.3

H. Daume. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007. 8.3.3, 8.4.2, 8.4.2

Department of the Army. *Police intelligence operations. Field Manual FM 3-19.50. Appendix D: Tactical Questioning*, July 2006. 1.4.2

J.-M. Dewaele and A. Furnham. Extraversion: the unloved variable in applied linguistic research. *Language Learning*, 49(3):509–544, 1999. 2.2.1, 4.1, 4.3, 4.3, 5.5.5

C. G. DeYoung, J. B. Peterson, and D. M. Higgins. Sources of openness/intellect: Cognitive and neuropsychological correlates of the fifth factor of personality. *Journal of Personality*, 73(4):825–858, 2005. 2.1.2

C. DiMarco and G. Hirst. A computational theory of goal-directed style in syntax. *Computational Liguistics*, 19(3):451–499, 1993. 2.4.3.2, 6.2

A. L. Edwards. The relationship between the judged desirability of a trait and the probability that it will be endorsed. *Journal of Applied Psychology*, 37:90–93, 1953. 3.7

M. Elhadad and J. Robin. An overview of SURGE: A reusable comprehensive syntactic realization component. Technical report, Department of Mathematics and Computer Science, Ben Gurion University, Beer Sheva, Israel, March 1996. 2.4

F. Enos, S. Benus, R. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg. Personality factors in human deception detection: Comparing human to machine performance. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2006. 1.4.1

H. J. Eysenck. Dimensions of personality: 16, 5 or 3? criteria for a taxonomic paradigm. *Personality and Individual Differences*, 12(8):773–790, 1991. 2.1.1

S. B. G. Eysenck, H. J. Eysenck, and P. Barrett. A revised version of the psychoticism scale. *Personality and Individual Differences*, 6(1):21–29, 1985. 2.1.2, 4.3, 4.4

L. A. Fast and D. C. Funder. Personality as manifest in word use: Correlations with self-report, acquaintance-report, and behavior. *Journal of Personality and Social Psychology*, in press, 2007. 2.2

C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998. 5.5.4

M. Fleischman and E. Hovy. Towards emotional variation in speech-based natural language generation. In *Proceedings of the 1st International Conference on Natural Language Generation*, pages 57–64, 2002. 2.4.3.1, 6.2

Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. In *Proceedings of the 15th International Conference on Machine Learning*, pages 170–178, 1998. 3.2.4

D. C. Funder. On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102:652–670, 1995. 3.7, 5.2

D. C. Funder. *The Personality Puzzle*. W. W. Norton & Company, New York, 2nd edition, 1997. 4.1.2

D. C. Funder and C. D. Sneed. Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology*, 64(3):479–490, 1993. 1.4

A. Furnham. Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*. Winley, 1990. 2.2.1, 4.1, 4.3, 4.3, 4.3, 4.3, 5.5.1, 5.5.2, 5.5.3

A. Furnham, C. J. Jackson, and T. Miller. Personality, learning style and work performance. *Personality and Individual Differences*, 27:1113–1122, 1999. 1.4.1, 1.4.2

A. Furnham and J. Mitchell. Personality, needs, social skills and academic achievement: A longitudinal study. *Personality and Individual Differences*, 12:1067–1073, 1991. 1.4.1

H. Giles and P. Powesland. *Speech style and social evaluation (European Monographs in Social Psychology)*. Harcourt Brace, New York, NY, 1975. 1.4.2

A. Gill and J. Oberlander. Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 456–461, 2003. 2.2.2, 3.2.2.1, 3.4.2, 4.1, 4.4, 4.4, 5.5.5

A. J. Gill and J. Oberlander. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368, 2002. 2.2.1, 3.2.2.1, 3.2.2.2, 7.2.2, 7.2.3

L. R. Goldberg. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229, 1990. 2.1.1, 3.7, 4.3, 6.2, 6.4, 7, 7.2.1

S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big five personality domains. *Journal of Research in Personality*, 37:504–528, 2003. 4.1.2, 6.2, 6.2, 6.4, 7.1, 8.1, 9.1, 9.2.1

M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar. Combining prosodic, lexical and cepstral systems for deceptive speech detection. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006. 1.4.1

S. J. Green and C. DiMarco. Stylistic decision-making in natural language generation. In *Proceedings of the 4th European Workshop on Natural Language Generation*, 1993. 2.4.3.2

F. Heylighen and J.-M. Dewaele. Variation in the contextuality of language: an empirical measure. *Context in Context, Special issue of Foundations of Science*, 7 (3):293–340, 2002. 1.3.2.1, 2.2.1, 3.3.3, 4.1, 4.3, 5.5.4

R. Higashinaka, M. A. Walker, and R. Prasad. An unsupervised method for learning generation lexicons for spoken dialogue systems by mining user reviews. *ACM Transactions on Speech and Language Processing*, 4(4), 2007. 5.5.2, 5.5.5, 5.6, 10.3

J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke. Distinguishing deceptive from non-deceptive speech. In *Proceedings of Interspeech - Eurospeech*, 2005. 1.4.1

R. Hogan, G. J. Curphy, and J. Hogan. What we know about leadership: Effectiveness and personality. *American Psychologist*, 49(6):493–504, 1994. 1.4.1

E. Hovy. *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum Associates, 1988. 2.4.3.1, 6.2, 9.3

M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2004. 5.5.2

R. C. Hubal, P. N. Kizakevich, C. I. Guinn, K. D. Merino, and S. L. West. The virtual standardized patient: Simulated patient-practitioner dialogue for patient interview training. In J. D. Westwood, H. M. Hoffman, G. T. Mogel, R. A. Robb, and D. Stredney, editors, *Envisioning Healing: Interactive Technology and the Patient-Practitioner Dialogue*. IOS Press, Amsterdam, 2000. 1.4.2

D. A. Infante. Teaching students to understand and control verbal aggression. *Communication Education*, 44(1):51–63, 1995. 4.1, 4.5

D. Z. Inkpen and G. Hirst. Near-synonym choice in natural language generation. In G. A. Nicolas Nicolov, Kalina Bontcheva and R. Mitkov, editors, *Recent Advances in Natural Language Processing III*. John Benjamins Publishing Company, 2004. 2.4.4.1

A. Isard, C. Brockmann, and J. Oberlander. Individuality and alignment in generated dialogues. In *Proceedings of the 4th International Natural Language Genera-*

tion Conference (INLG), pages 22–29, 2006. 2.4.4.1, 2.5, 6.2, 8.1, 2, 8.5, 10.1, 10.1

K. Isbister and C. Nass. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, 53(2):251 – 267, 2000. 6.4

O. P. John, E. M. Donahue, and R. L. Kentle. The "Big Five" Inventory: Versions 4a and 5b. Technical report, Berkeley: University of California, Institute of Personality and Social Research, 1991. 3.1, 3.2.1, 6.2, 6.4

O. P. John and S. Srivastava. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of personality theory and research*. New York: Guilford Press, 1999. 3.2.1, 4.1.2

K. Jokinen and K. Kanto. User expertise modelling and adaptivity in a speech-based e-mail system. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004. 2.3.2

M. Komarraju and S. J. Karau. The relationship between the Big Five personality traits and academic motivation. *Personality and Individual Differences*, 39:557–567, 2005. 1.4.1

W. Labov. Field methods of the project in linguistic change and variation. In J. Baugh and J. Sherzer, editors, *Language in Use*, pages 28–53. Prentice-Hall, 1984. 1.3.1, 1.3.2.1

I. Langkilde and K. Knight. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 704–710, 1998. 2.4.4.1

I. Langkilde-Geary. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 1st International Conference on Natural Language Generation*, 2002. 2.4, 2.4.4.1, 8, 8.1, 8.2, 8.3.2, 8.5, 9.3

B. Lavoie and O. Rambow. A fast and portable realizer for text generation systems. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 265–268, 1997. 5.4, 5.5.2, 5.5.6

D. Lin. Dependency-based evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*, 1998. 3.2.2.1

J. Liscombe, J. Venditti, and J. Hirschberg. Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Interspeech - Eurospeech*, 2003. 1.4.1

F. Mairesse and M. A. Walker. Learning to personalize spoken generation for dialogue systems. In *Proceedings of Interspeech - Eurospeech*, 2005. 2.3.1

F. Mairesse and M. A. Walker. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–503, 2007. 6.2, 9.3

F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500, 2007. 8.4, 10.1

P. Mallory and V. Miller. A possible basis for the association of voice characteristics and personality traits. *Speech Monograph*, 25:255–260, 1958. 3.2.2.3

W. C. Mann and S. A. Thompson. Rhetorical structure theory. Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988. 2.4.2, 4.2, 5.5.1

D. K. Marcus, S. O. Lilienfeld, J. F. Edens, and N. G. Poythress. Is antisocial personality disorder continuous or categorical? A taxometric analysis. *Psychological Medicine*, 36(11):1571–1582, 2006. 3.7, 6.4, 7

M. Mateas. The authoring bottleneck in creating AI-based interactive stories. In *Proceedings of the AAAI 2007 Fall Symposium on Intelligent Narrative Technologies*, 2007. 1.4.2, 10.2

M. Mateas and A. Stern. Façade: An experiment in building a fully-realized interactive drama. In *Proceedings of the Game Developers Conference, Game Design track*, 2003. 10.2, 10.1, 10.2

R. R. McCrae and P. T. Costa. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52:81–90, 1987. 4.1.2, 4.7

M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90:862–877, 2006. 2.2, 2.2.2, 3.2.1, 3.2.1, 3.2.2.1, 3.2.3, 3.2.3, 2, 3.9, 3.7, 4.1, 4.3, 4.3, 4.4, 4.4, 4.5, 4.5, 4.5, 4.5, 4.6, 4.6, 4.6, 4.6, 4.7, 4.7, 5.5.4, 5.5.5, 7.2.2, 8.4, 9.2.2, 10.1, 10.1, 10.3

I. A. Melčuk. *Dependency Syntax: Theory and Practice*. SUNY, Albany, New York, 1988. 5.5.2

G. Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, 2005. 1.4.1

C. Nass and K. Lee. Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181, 2001. 3.5.3

M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, 29:665–675, 2003. 1.4.1

W. T. Norman. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. *Journal of Abnormal and Social Psychology*, 66:574–583, 1963. 1.3.2.4, 2.1.1, 3.7, 4.1.2, 4.4, 6.4, 7

S. Nowson. *The Language of Weblogs: A study of genre and individual differences*. PhD thesis, University of Edinburgh, 2006. 4.1, 4.3, 4.5, 4.6, 4.7, 4.7, 4.7

S. Nowson and J. Oberlander. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *Proceedings of the International Conference on Weblogs and Social Media*, 2007. 2.3.3, 3.6

J. Oberlander and A. Gill. Individual differences and implicit language: personality, parts-of-speech, and pervasiveness. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 1035–1040, 2004a. 4.1, 4.3, 4.4, 4.4

J. Oberlander and A. Gill. Language generation and personality: two dimensions, two stages, two hemispheres? In *Proceedings from the AAAI Spring Symposium on Architectures for Modeling Emotion: Cross-Disciplinary Foundations*, pages 104–111, 2004b. 4.1, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 8.3.1

J. Oberlander and A. J. Gill. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 42: 239–270, 2006. 2.2, 2.2.1, 4.1

J. Oberlander and S. Nowson. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006. 1.4.1, 2.3.3, 3.2.4, 3.6, 7, 3.7

P.-Y. Oudeyer. Novel useful features and algorithms for the recognition of emotions in speech. In *Proceedings of the 1st International Conference on Speech Prosody*, pages 547–550, 2002. 1.4.1

D. S. Paiva. *Using Stylistic Parameters to control a natural language generation system*. PhD thesis, Information Technology Research Institute, University of Brighton, 2004. 8.3.2, 8.5

D. S. Paiva and R. Evans. Empirically-based control of natural language generation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 58–65, 2005. 2.4.4.2, 2.5, 6.2, 8.5, 9, 9.3

B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 115–124, 2005. 1.4.1

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002. 5.5.2

C. L. Paris. Tailoring object descriptions to a user's level of expertise. *Computational Linguistics*, 14(3):64–78, 1988. 2.3.2

S. V. Paunonen and D. N. Jackson. What is beyond the Big Five? plenty! *Journal of Personality*, 68(5):821–836, 2000. 2.1.1

D. Peabody and L. R. Goldberg. Some determinants of factor structures from personality-trait descriptor. *Journal of Personality and Social Psychology*, 57(3): 552–567, 1989. 2.1.1, 4.7

J. W. Pennebaker, M. E. Francis, and R. J. Booth. *Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum, Mahwah, NJ, 2001. 2.2.2, 3.4, 3.2.2.1, 3.2.2.1, 7.2.2

J. W. Pennebaker and L. A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77:1296–1312, 1999. 2.2, 2.2.1, 2.2.2, 3.2.1, 3.2.1, 3.2.2.1, 3.2.3, 3.2.3, 1, 2, 3.7, 3.6, 4.1, 4.1.1, 4.3, 4.3, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.6, 4.6, 4.6, 4.7, 4.7, 4.7, 4.7, 5.5.1, 5.5.2, 5.5.4, 7.2.2, 8.4, 9.2.2, 10.1, 10.1

J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54: 547–577, 2003. 2.2

M. J. Pickering and S. Garrod. Towards a mechanistic theory of dialogue. *Behavioral and Brain Sciences*, 27:169–225, 2004. 3.1

P. Piwek. A flexible pragmatics-driven language generator for animated agents. In *Proceedings of Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, 2003. 2.4.3.4

A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, 2005. 1.4.1

K. Porayska-Pomsta and C. Mellish. Modelling politeness in natural language generation. In *Proceedings of the 3rd International Conference on Natural Language Generation*, pages 141–150, 2004. 2.4.3.3, 6.2

B. Reeves and C. Nass. *The Media Equation*. University of Chicago Press, 1996. 1.4, 1.4.1, 1.4.2, 2.3.3, 3, 10.3

E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000. 1.2, 2.4, 2.4.2, 4, 4.2, 5.4, 5.5.4, 5.6

E. Rich. User modeling via stereotypes. *Cognitive Science*, 3:329–354, 1979. 2.3.1

R. Rienks and D. Heylen. Dominance detection in meetings using easily obtainable features. In H. Bourlard and S. Renals, editors, *Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, volume 3869 of *Lecture Notes in Computer Science*. Springer Verlag, 2006. 1.4.1

R. E. Riggio, C. Salinas, and J. Tucker. Personality and deception ability. *Personality and Individual Differences*, 9(1):189–191, 1988. 1.4.1

J. P. Rushton, H. G. Murray, and S. Erdle. Combining trait consistency and learning specificity approaches to personality, with illustrative data on faculty teaching performance. *Personality and Individual Differences*, 8:59–66, 1987. 1.4.1, 1.4.2

Z. Ruttkay, C. Dormann, and H. Noot. Embodied conversational agents on a common ground. In Z. Ruttkay and C. Pelachaud, editors, *From brows to trust: evaluating embodied conversational agents*, chapter 2, pages 27–66. Kluwer Academic Publishers, Norwell, MA, 2004. 1.4.2

R. E. Schapire. A brief introduction to boosting. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 2:1401–1406, 1999. 3.2.4

K. R. Scherer. Personality markers in speech. In K. R. Scherer and H. Giles, editors, *Social markers in speech*, pages 147–209. Cambridge University Press, 1979. 2.2, 2.2.1, 3.2.2.3, 3.4.3, 3.5.3, 4.1, 4.3, 4.3, 4.4, 10.3

K. R. Scherer. Vocal indicators of stress. In J. Darby, editor, *Speech evaluation in psychiatry*, pages 171–187. Grune & Stratton, New York, 1981. 4.1, 4.4, 4.4, 4.4, 5.5.4

K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256, 2003. 1.4.1

A. Siegman and B. Pope. Personality variables associated with productivity and verbal fluency in the initial interview. In *Proceedings of the 73rd Annual Conference of the American Psychological Association*, 1965. 2.2.1, 3.2.2.3, 4.1, 4.3, 4.3

A. W. Siegman. The telltale voice: Nonverbal messages of verbal communication. In S. Feldstein and A. W. Siegman, editors, *Nonverbal Behavior and Communication*, chapter 7, pages 183–243. Lawrence Erlbaum Associates, Hillsdale, NJ, 1978. 4.1, 4.4, 4.5, 4.7, 4.7

J. F. Sigurdsson. Computer experience, attitudes toward computers and personality characteristics in psychology undergraduates. *Personality and Individual Differences*, 12(6):617–624, 1991. 1.4.1

M. Slater, D.-P. Pertaub, C. Barker, and D. Clark. An experimental study on fear of public speaking using a virtual environment. In *3rd International Workshop on Virtual Rehabilitation*, 2004. 1.4.2

B. L. Smith, B. L. Brown, W. J. Strong, and A. C. Rencher. Effects of speech rate on personality perception. *Language and Speech*, 18:145–152, 1975. 2.2, 2.2.2, 3.2.2.3

A. Stent, R. Prasad, and M. A. Walker. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004. 2.4.4.1, 5.2, 5.1, 5.2, 5.5.3, 8.1

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26 (3):339–371, 2000. 3.7

A. Thorne. The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology*, 53:718–726, 1987. 4.1, 4.3, 5.5.1

P. D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 417–424, 2002. 1.4.1

K. Vogel and S. Vogel. L'interlangue et la personalité de l'apprenant. *International Journal of Applied Linguistics*, 24(1):48–68, 1986. 2.2, 3.2.2.2, 10.3

M. A. Walker, J. E. Cahn, and S. J. Whittaker. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the 1st Conference on Autonomous Agents*, pages 96–105, 1997. 2.4.3.3, 9.3

M. A. Walker, O. Rambow, and M. Rogati. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*, 16(3-4), 2002. 2.4.4.1, 8, 8.1, 8.2, 8.3.2, 8.5, 9.3

M. A. Walker, A. Stent, F. Mairesse, and R. Prasad. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)*, 30:413–456, 2007. 5.1, 5.2

M. A. Walker and S. Whittaker. Mixed initiative in dialogue: an investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 70–78, 1990. 3.2.2.2

M. A. Walker, S. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, and G. Vasireddy. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840, 2004. 2.3.1

N. Wang, W. L. Johnson, R. E. Mayer, P. Rizzo, E. Shaw, and H. Collins. The politeness effect: Pedagogical agents and learning gains. *Frontiers in Artificial Intelligence and Applications*, 125:686–693, 2005. 2.4.3.3

D. Watson and L. A. Clark. On traits and temperament: General and specific factors of emotional experience and their relation to the five factor model. *Journal of Personality*, 60(2):441–76, 1992. 1.3.2.4, 1.4.1, 4.4, 4.6, 4.7

J. B. Weaver. Personality and self-perceptions about communication. In J. C. McCroksey, J. A. Daly, M. M. Martin, and M. J. Beatty, editors, *Communication and Personality: Trait perspectives*, chapter 4, pages 95–118. Hampton Press, 1998. 4.1, 4.3, 4.3, 4.4

N. Webb, M. Hepple, and Y. Wilks. Error analysis of dialogue act classification. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue*, 2005. 3.7

J. Weizenbaum. Eliza–a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966. 2.4.1

J. Weizenbaum. *Computer power and human reason*. Freeman, San Francisco, 1976. 2.4.1

J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics*, 2005. 1.4.1

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistic*, 30(3):277–308, 2004. 1.4.1

J. Wilkie, M. A. Jacka, and P. J. Littlewood. System-initiated digressive proposals in automated human-computer telephone dialogues: the use of contrasting politeness strategies. *International Journal of Human-Computer Studies*, 62:41–71, 2005. 2.4.3.3

T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI)*, pages 761–769, 2004. 1.4.1

I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA, 2005. 3.2.4, 8.2, 9.1.2